Journal of Machine Learning Research 18 (2017) 1-45 Submitted 10/16; Revised 5/17; Published 12/17

Poisson Random Fields for Dynamic Feature Models

Valerio Perrone[*] V.PERRONE@WARWICK.AC.UK
Paul A. Jenkins[*†] P.JENKINS@WARWICK.AC.UK
Dario Spanò[*] D.SPANO@WARWICK.AC.UK
[*] *Department of Statistics*
[†] *Department of Computer Science*
University of Warwick
Coventry, CV4 7AL, UK

Yee Whye Teh Y.W.TEH@STATS.OX.AC.UK
Department of Statistics
University of Oxford
Oxford, OX1 3LB, UK

Editor: Lawrence Carin

Abstract

We present the *Wright-Fisher Indian buffet process* (WF-IBP), a probabilistic model for time-dependent data assumed to have been generated by an unknown number of latent features. This model is suitable as a prior in Bayesian nonparametric feature allocation models in which the features underlying the observed data exhibit a dependency structure over time. More specifically, we establish a new framework for generating dependent Indian buffet processes, where the Poisson random field model from population genetics is used as a way of constructing dependent beta processes. Inference in the model is complex, and we describe a sophisticated Markov Chain Monte Carlo algorithm for exact posterior simulation. We apply our construction to develop a nonparametric focused topic model for collections of time-stamped text documents and test it on the full corpus of NIPS papers published from 1987 to 2015.

Keywords: Bayesian nonparametrics, Indian buffet process, topic model, Markov chain Monte Carlo, Poisson random field

1. Introduction

The Indian buffet process (IBP) (Griffiths and Ghahramani, 2011) is a distribution for sampling binary matrices with any finite number of rows and an unbounded number of columns, such that rows are exchangeable while columns are independent. It is used as a prior in Bayesian nonparametric models where rows represent objects and columns represent an unbounded array of features. In many settings the prevalence of features exhibits some sort of dependency structure over time and modeling data via a set of independent IBPs may not be appropriate. There has been previous work dedicated to extending the IBP to dependent settings (e.g., Williamson et al., 2010a; Zhou et al., 2011; Miller et al., 2012; Gershman et al., 2015). In this paper we present a novel approach that achieves this by means of a particular time-evolving beta process, which has a number of desirable properties and is better-suited for a different range of applications.

For each discrete time $t_0, \ldots, t_T$ at which the data is observed, denote by Z_t the feature allocation matrix whose entries are binary random variables such that $Z_{ikt} = 1$ if object i possesses feature k at time t and 0 otherwise. Denote by $X_k(t)$ the probability that $Z_{ikt} = 1$, namely the probability that feature k is active at time t, and by $X(t)$ the collection of these probabilities at time t. The idea is to define a prior over the stochastic process $\{X(t)\}_{t\geq 0}$ which governs its evolution in continuous time. In particular, for each feature k, $X_k(t)$ evolves independently, while features are born and die over time. This is a desirable property in several applications such as in topic modeling, where at some point in time a new topic may be discovered (birth) or forsaken (death). Our model benefits from these properties while retaining a very simple prior where sample paths are continuous and Markovian. Finally, we show that our construction defines a time-dependent beta process from which the two-parameter generalization of the IBP is marginally recovered for every fixed time t (Ghahramani et al., 2007).

The stochastic process we use is a modification of the so-called Poisson Random Field (PRF), a model widely used in population genetics (e.g., Sawyer and Hartl, 1992; Hartl et al., 1994; Bustamante et al., 2001, 2003; Williamson et al., 2005; Boyko et al., 2008; Gutenkunst et al., 2009; Amei and Sawyer, 2010, 2012). In this setting, new features can arise over time and each of them evolves via an independent Wright-Fisher (W-F) diffusion. The PRF model describes the evolution of feature probabilities within the interval $[0, 1]$, allows for flexible boundary behaviours and gives access to several off-the-shelf results from population genetics about quantities of interest, such as the expected lifetime of features or the expected time feature probabilities spend in a given subset of $[0, 1]$ (see Ewens, 2004).

We apply the WF-IBP to a topic modeling setting, where a set of time-stamped documents is described using a collection of latent topics whose probabilities evolve over time. The WF-IBP prior allows us to incorporate time dependency into the focused topic model construction described in Williamson et al. (2010b), where the IBP is used as a prior on the topic allocation matrix determining which topics underlie each observed document. As opposed to several existing approaches to topic modeling, which require specifying the total number of topics in the corpus in advance (see for instance Blei et al., 2003), adopting a nonparametric approach saves expensive model selection procedures (such as the one described in Griffiths and Steyvers, 2004). This is also reasonable in view of the fact that the total number of topics in a corpus is expected to grow as new documents accrue. Most existing nonparametric approaches to topic modeling are not designed to capture the evolution of the popularity of topics over time and may thus not be suitable for corpora that span large time periods. On the other hand, existing nonparametric and time-dependent topic models are mostly based on the Hierarchical Dirichlet Process (HDP) (Teh et al., 2006), which implicitly assumes a coupling between the probability of topics and the proportion of words that topics explain within each document. This assumption is undesirable since rare topics may account for a large proportion of words in the few documents in which they appear. Our construction inherits from the static model presented in Williamson et al. (2010b) the advantage of eliminating this coupling. Moreover, it keeps inference straightforward while using an unbounded number of topics and flexibly capturing the evolution of their popularity continuously over time.

Section 2 introduces the beta process construction of the IBP and the PRF, providing the background for Section 3 where we modify the PRF to construct the WF-IBP, our

novel time-varying feature allocation model. Section 4 describes a fixed-K approximation and an intuitive inference scheme that is built upon in Section 5 to develop an exact MCMC algorithm for posterior inference with the WF-IBP. Section 6 combines the model with a linear-Gaussian likelihood and evaluates it on a range of synthetic data sets. Finally, Section 7 illustrates the application of the WF-IBP to topic modeling and presents results obtained on both synthetic data and on the real-world data set consisting of the full text of papers from the NIPS conferences between the years 1987 and 2015.

2. Background

Before introducing our time-varying feature allocation model we first review its two building blocks, namely the IBP and the PRF. We show how the beta process connects these two models, and we adjust the PRF to develop a time-dependent extension of the two-parameter IBP in the next section.

2.1 The beta and Indian buffet processes

A completely random measure B over a measurable space (Ω,Σ) is a random measure that assigns independent masses to disjoint subsets of Ω. Any positive completely random measure is uniquely characterized by a certain Levy measure on $\Omega \times \mathbb{R}^+$ (see Kingman, 1967). Denote by c a positive function over Ω (concentration function) and by B_0 a fixed measure on Ω (base measure), and assume the base measure B_0 is continuous. A beta process $\mathrm{B} \sim BP(c, B_0)$ on Ω is a completely random measure uniquely characterized by the Levy measure

$$\nu(d\omega, dx) = B_0(d\omega)c(\omega)x^{-1}(1-x)^{c(\omega)-1}dx,$$

where $x \in [0,1]$ and $\omega \in \Omega$. This completely random measure can be represented via a Poisson process. In order to draw $B \sim BP(c, B_0)$, draw a set of points $(\omega_i, x_i) \in \Omega \times [0,1]$ from a Poisson point process with base measure $\nu(d\omega, dx)$ and let

$$B = \sum_{k=1}^{\infty} x_k \delta_{\omega_k},$$

where $\{\omega_i\}$ are the atoms of the measure B and $\{x_i\}$ their respective weights. Given a realization B of a beta process $BP(c, \alpha B_0)$ with $\alpha > 0$, one can obtain a draw from the IBP (Thibaux and Jordan, 2007). This is done by sampling each row z_i, $i = 1, \ldots, N$, of an allocation matrix Z from a Bernoulli process $z_i \mid B \sim BeP(B)$ defined as

$$z_i = \sum_{k=1}^{\infty} p_{ik}\delta_{\omega_k}, \quad p_{ik} \overset{iid}{\sim} \text{Bernoulli}(x_k).$$

Notice that as the location of the atoms is not in fact relevant for constructing the matrix Z, the IBP depends only on the so-called mass parameter $B_0(\Omega) = \alpha$ and on $c(\omega)$. In the standard IBP, $c(\omega)$ is set to be identically 1, so that the Levy measure of the corresponding one-parameter beta process is

$$\nu(d\omega, dx) = B_0(d\omega)x^{-1}dx. \tag{1}$$

If $c(\omega)$ is equal to a constant β we have the two-parameter generalization of the IBP (Thibaux and Jordan, 2007).

2.2 The Wright-Fisher model

The starting point of the PRF is the Wright-Fisher (W-F) model from population genetics (Ewens, 2004), which we briefly summarize here. Consider a finite population of organisms of size G such that $i \in \{0, 1, \ldots, G\}$ individuals have the mutant version of a gene at generation k, while the rest has the non-mutant variant. Assume that each individual produces an infinite number of gametes such that the gametes yielded by a non-mutant become mutant with probability μ_G and, conversely, those yielded by a mutant become non-mutant with probability β_G. Finally, assume that the next generation of G individuals is formed by simple random sampling from this infinite pool of gametes. The evolution of the number $Y^G(k)$ of mutant genes at time k is described by a Markov chain on the discrete space $\{0, \ldots, G\}$. The transition probability p_{ij} of switching from i mutants (at time k) to j mutants (at time $k+1$) is given by the following binomial sampling formula:

$$p_{ij} = \binom{G}{j} (\Psi_i)^j (1 - \Psi_i)^{G-j},$$
$$\Psi_i = \frac{i(1-\beta_G) + (G-i)\mu_G}{G}.$$

Assume the initial state is $Y^G(0) = y_0$ and denote the resulting Markov chain by $Y^G = (Y^G(k))_{k=1,2,\ldots} \sim \text{W-F}^G(\mu_G, \beta_G)$. Notice that, if $\mu_G = 0$ and/or $\beta_G = 0$, the states 0 and/or G are absorbing states that respectively correspond to the extinction and fixation of the mutation.

A continuous-time diffusion limit of the W-F model can be obtained, by rescaling time as $t = k/G$, and taking $G \to \infty$. The Markov chain $Y^G(\lfloor Gt \rfloor)/G$ converges to a diffusion process on $[0, 1]$ (see Ethier and Kurtz, 1986; Sawyer and Hartl, 1992) which obeys the one-dimensional stochastic differential equation

$$dX(t) = \gamma(X(t))dt + \sigma(X(t))dB(t),$$

where

$$\gamma(x) = \frac{1}{2}[\mu(1-x) - \beta x], \tag{2}$$
$$\sigma(x) = \sqrt{x(1-x)}, \tag{3}$$

with some initial state $X(0) = x_0$, over the time interval $t \in [0, T]$, with rescaled parameters $\mu = \lim_{G\to\infty} 2G\mu_G$, $\beta = \lim_{G\to\infty} 2G\beta_G$, and with $B(t)$ denoting a standard Brownian motion. The terms $\gamma(x)$ and $\sigma(x)$ are respectively referred to as the drift term and the diffusion term. Denote the diffusion process as $X \sim \text{W-F}(\mu, \beta)$.

When $x(t) \to 0$ (respectively $x(t) \to 1$), then the diffusion term tends to 0 while the drift term tends to $\frac{\mu}{2}$ (respectively $-\frac{\beta}{2}$), preventing absorption at 0 or 1 provided that $\mu > 0$ (respectively $\beta > 0$). Otherwise, 0 is an absorbing extinction state (respectively, 1 is an absorbing fixation state). Moreover, if both $\mu, \beta > 0$ then the diffusion is ergodic and has a stationary distribution that is a Beta(μ, β).

As there exists no closed form expression for its transition function, simulating from the W-F diffusion requires non-trivial computational techniques. The method we used is outlined in Dangerfield et al. (2012), a stochastic Taylor scheme tailored to the W-F diffusion. A novel alternative approach that allows for exact simulation has very recently been proposed by Jenkins and Spanò (2017). Finally, note that simulating from the W-F diffusion implies doing so for a given diffusion time unit Δt. While in population genetics diffusion time is related to the population size, in different applications it can be used to regulate the degree of sample path variance, which is linear in time and equal to $\mathrm{Var}(x(1-x)\Delta t)$.

2.3 The Poisson random field

The W-F model describes the evolution of a gene at one particular site. The PRF generalizes it to modeling an infinite collection of sites, each of which evolves independently according to the W-F model. As before, we start with a model with a population of finite size G, before taking the diffusion limit as $G \to \infty$. For a site i at which some individuals carry the mutant gene, denote by $X_i(k)$ the fraction of mutants in generation k. Each site evolves independently according to the W-F$^G(0,0)$ model. Further suppose that at each generation k a number of mutations $M \sim \mathrm{Poisson}(\nu_G)$ arise in new sites with indices $j_1, j_2, \ldots j_M$. ν_G is referred to as the immigration parameter of the PRF. Assume that each of the new mutations occurs at a new site in a single individual, with initial frequency $X_{j_m}(k) = 1/G$. Subsequently, each new process $X_{j_m}(k+1), X_{j_m}(k+2), \ldots$ evolves independently according to the W-F$^G(0,0)$ model as well (Figure 1). As with pre-existing mutant sites, each process eventually hits one of the boundaries $\{0,1\}$ and stays there (we say that the mutation is extinct/has been fixed).

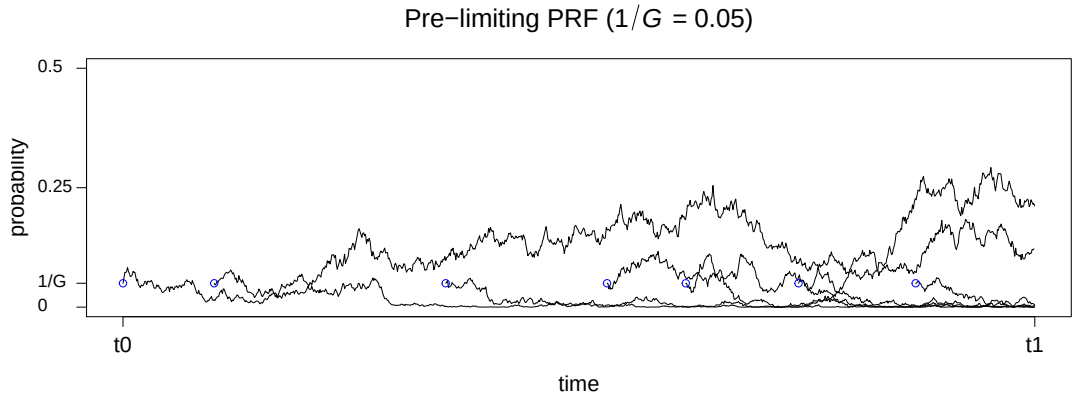


Figure 1: Evolution of mutant sites over time in the pre-limiting PRF model. The blue circles indicate mutations arising at a new site.

Consider the limit $G \to \infty$, so that after the same rescaling $t = k/G$ of time as in Section 2.2 each site evolves as an independent W-F diffusion $X_i \sim \text{W-F}(0,0)$. We also assume that $\nu_G \to \alpha$ as $G \to \infty$. This means that in the diffusion time scale the immigration rate is $G\nu_G \to \infty$, which suggests that the number of sites with mutant genes should explode. However, the initial frequency of each diffusion is $1/G \to 0$ as $G \to \infty$, and moreover 0 is an absorbing state. It can be shown (Sawyer and Hartl, 1992; Amei and Sawyer, 2010) that only

$O(G^{-1})$ of the newborn processes are not almost immediately absorbed. Therefore, there is a balance between the infinite number of newborn mutations and the infinite number of them going extinct in the first few generations, in such a way that the net immigration rate is $O(G\nu_G \times G^{-1}) = O(\alpha)$, and hence the limiting stationary measure is nontrivial. Provided that we remove from the model all sites whose frequency hits either the boundary 1 or 0, Sawyer and Hartl (1992) prove that the limiting distribution of the fractions of mutants in the interval $[0, 1]$ is a Poisson random field with mean density

$$\alpha x^{-1} dx. \tag{4}$$

Interestingly, the rate measure of the PRF coincides with the distribution of weights in the one-parameter beta process given in Equation (1). This means that at equilibrium the number of sites whose frequencies $X_i(t)$ are in any given interval $(a, b]$ is Poisson distributed with rate $\alpha \int_a^b x^{-1} dx$, and these are independent for nonoverlapping intervals. Integrating (4) over $[0, 1]$ shows that the number of mutations in the population that has not been fixed or gone extinct is infinite. However, most mutations are present in a very small proportion of the population.

3. Time-Varying Feature Allocation Model

The derivation of the PRF in the previous section shows that, as long as sites reaching frequency 1 or 0 are removed from the model, the equilibrium distribution of the PRF is related to the one-parameter beta process. In this section we generalize the PRF so that it is better adapted to applications in feature allocation modeling. Specifically, we identify mutant sites with features, and identify the proportion of the population having the mutant gene with the probability of the feature occurring in a data observation. The PRF can be then used in a time-varying feature allocation model whereby features arise at some unknown time point, change their probability smoothly according to a W-F diffusion process and eventually die when their probability reaches zero.

3.1 The WF-IBP

Recall from the previous section that mutant sites whose frequency hits 1 are removed from the PRF model. This means that features with high probability of occurrence can be removed from the model instantaneously, which does not make modeling sense. Instead, one expects a feature probability to change smoothly and to be removed from the model only once its probability of occurrence is *small*. A simple solution to this conundrum is to prevent 1 from being an absorbing state by using instead a W-F$(0, \beta)$ diffusion with $\beta > 0$. This is a departure from Sawyer and Hartl (1992), due to the differing modeling requirements of genetics versus feature allocation modeling. At the same time, both models let features disappear once their probability gets to 0, which is suitable from a feature allocation perspective and, as we now see, allows for a nontrival equilibrium mean density.

We shall denote the modified stochastic process as PRF(α, β). The following theorem derives the equilibrium mean density of PRF(α, β), with proof given in Appendix A:

Theorem 1 *The equilibrium mean density of the* PRF(α, β) *is*

$$l(x) = \alpha x^{-1}(1 - x)^{\beta - 1} dx. \tag{5}$$

In other words, the mean density of the PRF(α, β) is the Lévy measure of the two-parameter beta process, with the immigration rate α identified with the mass parameter, and β identified with the concentration parameter of the beta process. When $\beta = 1$ the one-parameter beta process is recovered. We assume that the initial distribution of PRF(α, β) is its equilibrium distribution, that is, a Poisson random field with mean density (5), so that the marginal distribution of the PRF at any point in time is the same.

We will now make the connection more precise by specifying how a PRF can be used in a time-varying feature allocation model. Denote by $X_k(t)$ the probability of feature k being active at time t and define our PRF as the stochastic process $X := \{X_k(t)\}$. Assume that at a finite number of time points $t = t_0, \ldots, t_T$ there are N_t objects whose observable properties depend on a potentially infinite number of latent features. Let D_{it} be the observation associated with object $i = 1, \ldots, N_t$ at time $t = t_0, \ldots, t_T$. Consider a set of random feature allocation matrices Z_t such that entry Z_{ikt} is equal to 1 if object i at time t possesses feature k, and 0 otherwise. Let $Z := \{Z_{ikt}\}$. Finally, let ρ_k be some latent parameters of feature k and $\rho = \{\rho_k\}$ be the set of all feature parameters. Our complete WF-IBP model is given as follows.

$$\begin{aligned} X &\sim \text{PRF}(\alpha, \beta), \\ Z_{ikt} \mid X &\overset{ind}{\sim} \text{Bernoulli}(X_k(t)), \\ \rho_k &\overset{iid}{\sim} H, \\ D_{it} \mid \rho, Z_{it} &\overset{ind}{\sim} F(\{\rho_k : Z_{ikt} = 1\}), \end{aligned} \tag{6}$$

where $i = 1, \ldots, N_t$, $t = t_0, \ldots, t_T$ and $k = 1, 2, \ldots, H$ is the prior distribution for feature parameters, and where $F(\rho)$ is the observation model for an object with a set of features with parameters ρ.

Since the feature probabilities X have marginal density (5), at each time t the feature allocation matrix Z_t has marginal distribution given by the two-parameter Indian buffet process (Thibaux and Jordan, 2007). Further, since X varies over time, the complete model is a time-varying Indian buffet process feature allocation model. The corresponding De Finetti measure would then be a time-varying beta process. More precisely, this is the measure-valued stochastic process $G = \{G(t)\}$ where

$$G(t) = \sum_{k=1}^{\infty} X_k(t) \delta_{\rho_k},$$

which has marginal distribution given by a beta process with parameters α, β and base distribution H. We denote the distribution of G as WFBP(α, β, H). We can also express the feature allocations using random measures as well. In particular, let

$$B_{it} = \sum_{k=1}^{\infty} Z_{ikt} \delta_{\rho_k}$$

be a Bernoulli process BeP$(G(t))$ with mean measure given by the beta process $G(t)$ at time t. An equivalent way to express our model (6) using the introduced random measures

is then

$$\begin{aligned} G &\sim \mathrm{WFBP}(\alpha, \beta, H), \\ B_{it} \mid G &\sim \mathrm{BeP}(G(t)), \\ D_{it} \mid B_{it} &\sim F(B_{it}), \end{aligned}$$

where WFBP denotes our time-varying beta process, and we have used $F(B)$ to denote the same observation model as before, but with B being a random measure with an atom for each feature, and whose location is the corresponding feature parameter. In the following, we use the notation introduced in (6) instead of in terms of beta and Bernoulli processes for simplicity.

As in the two-parameter IBP, α and β separately control the distribution of the number of features per object and the total number of features. In addition, the discrete time points $t_0, \ldots, t_T$ at which the observations are given influence the number of time units for which the W-F diffusions should be simulated. This introduces a parameter that regulates the strength of the time-dependency or accounts for gaps of varying sizes between successive observations. The time parameter can also be chosen according to the expected life time of features (see Chapter 15 of Karlin and Taylor, 1981). A detailed description of how to simulate from the model is given in Appendix B.

3.2 Related models

The WF-IBP fits a line of research that aims at introducing dependency structures into the IBP. A number of these extensions are designed to drop the exchangeability assumption from the IBP by coupling the rows and columns of the feature allocation matrix, and are thus orthogonal to our work (see Zhou et al., 2011; Miller et al., 2012; Gershman et al., 2015). What we are rather interested in achieving is partial exchangeabilty, whereby objects can be permuted independently at each time point without changing the probability of the process. This is necessary for time-dependent topic models where each time has a different set of documents and there is no correspondence between documents at different times.

One example that is more closely related to our model is the dependent Indian buffet process (dIBP) (Williamson et al., 2010a), which can also be applied to the partially exchangeable case. The dIBP uses a hierarchical Gaussian process to introduce couplings between features and items in such a way that, for an appropriate choice of the kernel, items can be permuted independently at each time. Although both the dIBP and the WF-IBP try to achieve the same goal, their methodologies are substantially different: while the former introduces dependencies at the feature-matrix level, the latter does so at the beta-process level. The construction of a dependent beta process represents a significantly different research direction, and is indeed anticipated as future work in Williamson et al. (2010a). First, an important consequence is that the WF-IBP prior describes the evolution of feature probabilities explicitly, whereas the dIBP fixes them and effectively describes a time-evolving Bernoulli process. The WF-IBP is then preferable in all settings where one is interested in a direct interpretation of the evolution of features. Second, the dIBP suffers from a less flexible boundary behaviour as it does not allow features to be born and die over time. Third, the dIBP is based on the stick-breaking construction of the IBP, which is only available for the one-parameter IBP; instead, our approach extends the two-parameter

IBP and models the dimensionality of the feature allocation matrix and its sparsity independently.

4. Fixed-K approximation

In order to give more intuition on the exact inference with the WF-IBP that is developed in Section 5, we first describe a finite-dimensional approximation where the number of features is a finite number K, and show that this marginally converges to the WF-IBP as $K \to \infty$.

Assume that the random feature allocation matrix Z_t has a fixed number of features, say K. First, let $\alpha, \beta > 0$ and consider the beta-binomial model

$$\begin{aligned} Z_{ikt} \mid \{X_k(t) = x_k(t)\} &\overset{iid}{\sim} \text{Bernoulli}(x_k(t)), \\ X_k(t) &\overset{iid}{\sim} \text{Beta}\left(\frac{\alpha\beta}{K}, \beta\right), \end{aligned}$$

$\forall k = 1, \ldots, K, \forall i = 1, \ldots, N_t$. This coincides with the pre-limiting model of the two-parameter IBP presented in Ghahramani et al. (2007). Then, for each feature, think of the $\text{Beta}\left(\frac{\alpha\beta}{K}, \beta\right)$ distribution as the stationary distribution of a W-F diffusion with parameters $\frac{\alpha\beta}{K} > 0$ and $\beta > 0$. This suggests making the model time-dependent by letting each feature evolve, starting at stationarity, as an independent W-F diffusion with these parameters. Generate for all times $t = t_0, \ldots, t_T$ the binary variables z_{ikt} as

$$\begin{aligned} Z_{ikt} \mid \{X_k(t) = x_k(t)\} &\overset{iid}{\sim} \text{Bernoulli}(x_k(t)), \\ X_k &\sim \text{WF}\left(\frac{\alpha\beta}{K}, \beta\right), \end{aligned}$$

$\forall k = 1, \ldots, K, \forall i = 1, \ldots, N_t$. In this way, the closer two time points, the stronger the dependency between the probabilities of a given feature (Figure 2). Moreover, as we assume the W-F diffusion to start at stationarity, this construction coincides marginally with the beta-binomial model. The parameters of the W-F diffusion are positive, so that neither fixation nor absorption ever occurs and the number K of features remains constant.

4.1 Fixed-K MCMC inference

Given a set of observations D, a natural inference problem would be to recover the latent feature allocation matrices $Z = \{Z_t\}_{t=t_0}^{t_T}$ responsible for generating the observed data, the underlying feature probabilities X and their parameters ρ. Inference is straightforward; we propose the following updates.

- $Z \mid X, D, \rho$ via Gibbs sampling.
- $\rho \mid D, Z$ according to the likelihood model.
- $X \mid Z$ via Particle Gibbs.

Consider first the Gibbs sampling step to perform posterior inference over the matrices Z. Denote by $Z_{-(ik)t}$ all the components of the matrix Z_t excluding Z_{ikt}, and by Z_{i-kt} all

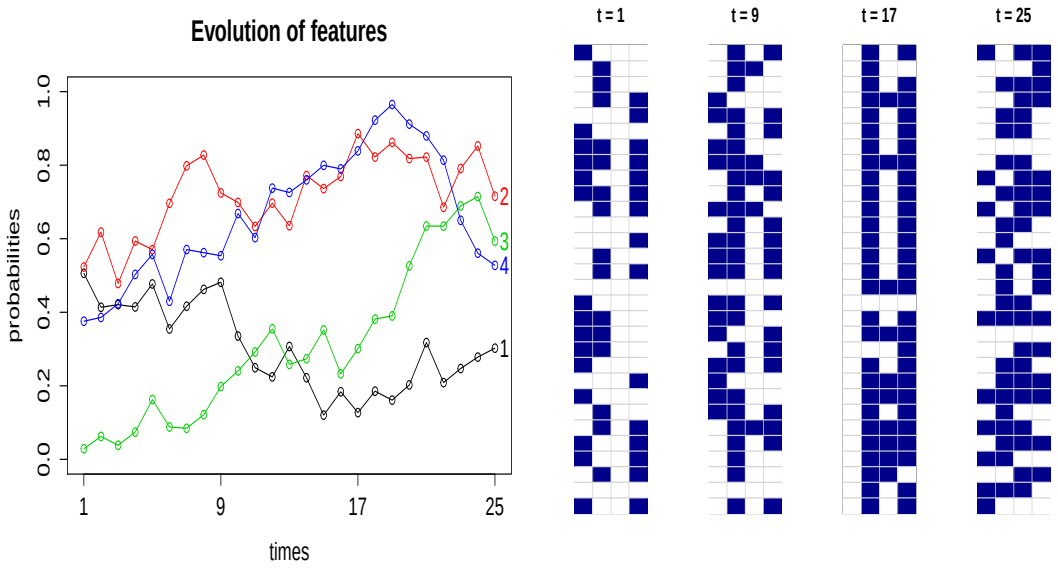


Figure 2: Left: underlying feature probabilities over time. Right: corresponding feature allocation matrices. Rows represent objects and columns features, which can be active (blue) or inactive (white).

the components in row i excluding k. We can easily derive for all t the distribution of a given component Z_{ikt} conditioning on the state of all other components $Z_{-(ik)t}$, on data D_{it}, on the parameters ρ and on the prior probability $X_k(t)$ of feature k. The full conditional probability of the entry Z_{ikt} being active is equal to

$$P(Z_{ikt} = 1 \mid Z_{-(ik)t}, X_k(t), D_{it}, \rho) \propto x_k(t) P(D_{it} \mid Z_{i-kt}, Z_{ikt} = 1, \rho). \tag{7}$$

By the same token, the full conditional probability of the entry Z_{ikt} being inactive is

$$P(Z_{ikt} = 0 \mid Z_{-(ik)t}, X_k(t), D_{it}, \rho) \propto (1 - x_k(t)) P(D_{it} \mid Z_{i-kt}, Z_{ikt} = 0, \rho). \tag{8}$$

As the matrices Z are conditionally independent given the feature probabilities X, equations (7) and (8) can be used to sample the matrices Z independently given the respective feature probabilities at each time. Note that the likelihood $P(D_{it} \mid Z_t, \rho)$ needs to be specified according to the problem at hand. A typical choice, detailed in Section 6, is the linear-Gaussian likelihood model, whose parameters can easily be integrated out (Griffiths and Ghahramani, 2011). The update $\rho \mid D, Z$ over the feature parameters is also specific to the likelihood model and, as we will illustrate, can easily be derived in conjugate models such as the linear-Gaussian one.

Consider now the Particle Gibbs (PG) step (Andrieu et al., 2010) to perform Bayesian inference on the feature trajectories continuously over the interval $[t_0, t_T]$. As the prior probability of each feature is a Beta $\left(\frac{\alpha\beta}{K}, \beta\right)$ and the column-wise sums of Z_{t_0} are realizations from binomial distributions, by conjugacy we have

$$X_k(t_0) \mid \{Z_{t_0} = z_{t_0}\} \sim \text{Beta}\left(\frac{\alpha\beta}{K} + n_{kt_0}, \beta + N_{t_0} - n_{kt_0}\right),$$

where $k = 1, \ldots, K$, and $n_{kt} := \sum_{i=1}^{N_t} z_{ikt}$ denotes the number of objects in matrix Z_t possessing feature k. This posterior distribution can be therefore used to draw the whole set of features at time t_0 and the trajectories in the interval $[t_0, t_T]$ can be obtained via PG.

More precisely, start with an initial reference trajectory $x^{r,k}_{t_0:t_T} := (x^{r,k}_{t_0}, \ldots, x^{r,k}_{t_T})$ for $k = 1, \ldots, K$ and, independently for each feature, iterate the following procedure. Draw a given number of particles from the posterior beta distribution at time t_0 and propagate them forward to time t_1 according to $\mathrm{WF}(\frac{\alpha\beta}{K}, \beta)$. At time t_1, assign each of these particles and the reference feature a weight given by the binomial likelihood of seeing that feature active in $n(t)$ objects out of $N(t)$ in $Z(t)$, i.e., $x_k(t)^{n_{kt}}(1 - x_k(t))^{N_t - n_{kt}}$. Sample the weighted particles with replacement and propagate the off-springs forward. This corresponds to using a bootstrap filter with multinomial resampling, but other choices to improve on the performance of the sampler can be made (see Andrieu et al., 2010). Repeat this procedure up to time t_T and sample only one particle at that time. Reject all the others and keep the trajectory that led to the sampled particle as the reference trajectory for the next iteration. Notice that the reference feature is kept intact throughout each iteration of the algorithm.

This procedure is illustrated more precisely by Algorithm 1, which needs to be iterated independently for each feature to provide posterior samples from their trajectories. To simplify the notation, we drop the index k and write $x^i_t | x^i_{t_0} \sim \mathrm{WF}(\frac{\alpha\beta}{K}, \beta)$ for $t \in [t_0, t_1]$ to denote the following: simulate from a W-F diffusion with initial value $x(t_0)$ and set $X(t_1) = x(t_1)$, the value of the diffusion at time t_1.

4.2 Approximation for large K

As already noted, the marginal distribution with the fixed-K approximation corresponds to the beta-binomial model, which is the pre-limiting model of the two-parameter IBP. As a consequence, at any fixed time t and as $K \to \infty$, the fixed-K approximation converges to the two-parameter generalization of the IBP, which in turns coincides with the marginal distribution of the WF-IBP. An aspect of interest is then whether the whole dynamics of the fixed-K approximation can be used as a finite approximation of the infinite model in such a way that, the larger K, the better the approximation. Two caveats need to be noted. First, only in the infinite model can features be born. For large K, however, the number of particles in the fixed-K approximation whose mass is close to zero becomes so large that, with a sufficient amount of time, some of them gain enough mass to become 'visible'. The behavior of these particles resembles the behavior of the newborn features of the infinite model. Second, as the fixed-K approximation has an upwards drift equal to $\alpha\beta/K$, only the infinite model allows for features to be absorbed at 0. This discrepancy is however overcome by the fact that, when K goes to infinity, 0 behaves like an absorbing boundary, in that features get trapped at probabilities close to 0. For these reasons, a comparison of the two models requires relabeling the particles in the finite model in such a way that, whenever a particle goes below a certain threshold $\epsilon \approx 0$, it is considered to have gone extinct, while if its probability is below ϵ and later exceeds ϵ the particle is labeled as newborn. We choose this threshold to be $\epsilon = 1/K$, as then $\lim_{K\to\infty} \epsilon = 0$.

Taking these caveats into account, we performed an empirical comparison of the fixed-K approximation with the infinite model. Consider the joint distribution at two given time points $t_0 = 0$ and $t_1 = 1$ of the feature probabilities, first in the fixed-K and then in the

Algorithm 1: Particle Gibbs

Input: Reference trajectory $x^r_{t_0:t_T}$; M.

Set $x^M_{t_0} = x^r_{t_0}$;

Draw $x^i_{t_0} \sim \text{Beta}(\frac{\alpha\beta}{K} + n_{t_0},\, \beta + N_{t_0} - n_{t_0})$ for $i = 1, \ldots, M-1$;

Simulate $x^i_t | x^i_{t_0} \sim \text{WF}(\frac{\alpha\beta}{K}, \beta)$ for $t \in [t_0, t_1]$ for $i = 1, \ldots, M-1$;

Set $x^M_{t_1} = x^r_{t_1}$;

Compute $w^i_{t_1} = (x^i_{t_1})^{n_{t_1}} (1 - x^i_{t_1})^{N_{t_1} - n_{t_1}}$ for $i = 1, \ldots, M$;

Sample $\bar{x}^i_{t_1}$ with $P(\bar{x}^i_{t_1} = x^i_{t_1}) \propto w^i_{t_1}$ for $i = 1, \ldots, M-1$;

Set $\bar{x}^M_{t_1} = x^r_{t_1}$;

Simulate $x^i_t | x^i_{t_1} \sim \text{WF}(\frac{\alpha\beta}{K}, \beta)$ for $t \in [t_1, t_2]$ for $i = 1, \ldots, M-1$;

Set $j \leftarrow 2$;

while $t_j < t_T$ **do**

  Set $x^M_{t_j} = x^r_{t_j}$;

  Compute $w^i_{t_j} = (x^i_{t_j})^{n_{t_j}} (1 - x^i_{t_j})^{N_{t_j} - n_{t_j}} w^i_{t_{j-1}}$ for $i = 1, \ldots, M$;

  Sample $\bar{x}^i_{t_j}$ with $P(\bar{x}^i_{t_j} = x^i_{t_j}) \propto w^i_{t_j}$ for $i = 1, \ldots, M-1$;

  Set $\bar{x}^M_{t_j} = x^r_{t_j}$;

  Simulate $x^i_t | \bar{x}^i_{t_j} \sim \text{WF}(\frac{\alpha\beta}{K}, \beta)$ for $t \in [t_j, t_{j+1}]$ for $i = 1, \ldots, M-1$;

  Set $j \leftarrow j + 1$;

end

Compute $w^i_{t_T} = (x^i_{t_T})^{n_{t_T}} (1 - x^i_{t_T})^{N_{t_T} - n_{t_T}} w^i_{t_{T-1}}$ for $i = 1, \ldots, M$;

Sample r_{new} with $P(r_{new} = i) \propto w^i_{t_T}$, where $i = 1, \ldots, M$;

Output: New reference trajectory $x^{r_{new}}_{t_0:t_T}$.

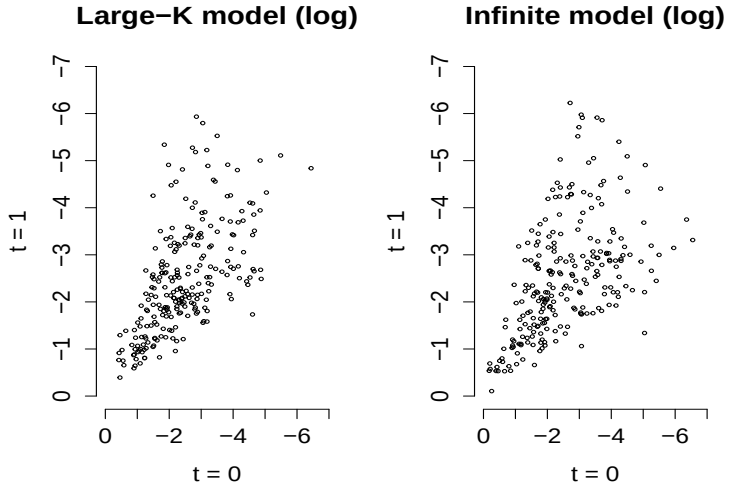


Figure 3: Scatterplot of the log feature probabilities greater than $1/K$ at time $t_0 = 0$ and at time $t_1 = 1$ to compare the fixed-K ($K = 1000$) and infinite model.

infinite model. Separately for each model, we took 1000 samples of the feature probabilities at time 0 and at time 1, excluding the ones below $1/K$. Figure 3 shows the logarithm of these values for the two models, suggesting a remarkable similarity between the two underlying joint distributions. The validity of this comparison is supported by the maximum mean discrepancy (mmd) test (Gretton et al., 2006), which does not reject the null hypothesis of the two joint distributions being the same. Although this suggests a strong similarity between the dynamics of the fixed-K approximation and the infinite model, we leave a proof of the convergence of these joint distributions as $K \to \infty$ for future work.

5. Exact MCMC inference

Building on the fixed-K approximation, we now develop a sophisticated MCMC algorithm for exact inference with the WF-IBP. The first point to notice is that, while in the fixed-K approximation the total number of features is constant over time and equal to K, in the WF-IBP model this is not a finite number. In order to use the WF-IBP for inference it is necessary to augment the state space with the features that are not seen in the feature allocation matrices, but simulating the dynamics of the PRF would require generating an infinite number of features, which is clearly unfeasible. One way to deal with this could be to resort to some sort of truncation, considering only features whose probability is greater than a given threshold and are likely to be seen in the data. We rather choose this truncation level adaptively by introducing a collection of slice variables $\{S_t\}_{t=1}^{T_t}$ and adopting conditional slice sampling (Walker, 2007; Teh et al., 2007). This scheme, which is detailed in Section 5.1, has the advantage of making inference tractable without introducing approximations.

Partition the set of features into two subsets, one containing the features that have been seen at least once among times $t = t_0, \ldots, t_T$, and the other containing the features that have never been seen for all $t = t_0, \ldots, t_T$, so that $X = X_{\text{seen}} \cup X_{\text{unseen}}$. Since the unseen

features cannot be identified individually based on the matrices Z, and as all features are conditionally independent given Z, we consider seen and unseen features separately. As for the seen features, we use the same Particle Gibbs scheme as in the fixed-K approximation. As for the unseen features, we simulate them via a thinning scheme. The exact MCMC inference scheme can be then summarized by the following updates.

- $Z \mid X, S, D, \rho$ via Gibbs sampling.
- $S \mid Z, X$ via slice sampling.
- $\rho \mid D, Z$ according to the likelihood model.
- $X_{\text{seen}} \mid Z$ via Particle Gibbs.
- $X_{\text{unseen}} \mid S$ via thinning.

We present each of these steps in Sections 5.1 and 5.2, which build on the simpler inference scheme developed in Section 4.

5.1 Gibbs and slice sampling

The first step is to augment the parameter space with a set of slice variables. Given the feature allocation matrices Z_t at times $t = t_0, \dots, t_T$, draw a slice variable $S_t \sim \text{Uniform}[0, x^{min}(Z_t)]$ for each time t, where $x^{min}(Z_t)$ is the minimum among the probabilities of the features seen in the feature matrix Z_t. In this way, when conditioning on the value s_t of the slice variable, we have a truncation level s_t and only need to sample the finite number of features whose probability is above this threshold (Teh et al., 2007). In other words, for all $t = t_0, \dots, t_T$, we only need to update the columns of Z_t whose corresponding feature probability $x_k(t)$ is greater than or equal to the slice variable s_t (note that these include both seen and currently unseen features). Observe that we defined a different slice variable for each time point, while an alternative choice could have been drawing a single slice from a uniform between 0 and the minimum feature probability across all times. Having multiple slice variables makes it possible to simulate fewer feature trajectories while keeping inference exact, reducing the computational cost of simulating features with small probabilities of being active. Although this comes at the cost of a larger number of parameters, in experiments we find that having multiple slices does not compromise mixing nor predictive performance.

Accounting for the slice variables, the full conditional probability of the entry Z_{ikt} being active is directly proportional to

$$\begin{aligned} &P(Z_{ikt} = 1 \mid Z_{-(ik)t}, X_k(t), D_{it}, \rho) P(S_t \mid Z_{ikt} = 1, Z_{-(ik)t}) \propto \\ &\quad x_k(t) P(D_{it} \mid Z_{i-kt}, Z_{ikt} = 1, \rho) \frac{1[0 \leq s_t \leq x^{min}(Z_{-(ik)t}, Z_{ikt} = 1)]}{x^{min}(Z_{-(ik)t}, Z_{ikt} = 1)}, \end{aligned} \tag{9}$$

while the full conditional probability of the entry Z_{ikt} being inactive is directly proportional to

$$\begin{aligned} &P(Z_{ikt} = 0 \mid Z_{-(ik)t}, X_k(t), D_{it}, \rho) P(S_t \mid Z_{ikt} = 0, Z_{-(ik)t}) \propto \\ &\quad (1 - x_k(t)) P(D_{it} \mid Z_{i-kt}, Z_{ikt} = 0, \rho) \frac{1[0 \leq s_t \leq x^{min}(Z_{-(ik)t}, Z_{ikt} = 0)]}{x^{min}(Z_{-(ik)t}, Z_{ikt} = 0)}. \end{aligned} \tag{10}$$

The term $P(S_t \mid Z_t)$ is not constant as updating Z_{ikt} for a currently unseen feature may modify the value of the minimum probability of the active features.

5.2 Particle Gibbs and thinning

Assume that the feature allocation matrices Z_t are given at the time points $t = t_0, \ldots, t_T$ and we are interested in inferring the probabilities X of the underlying features. This section gives the details of inference for each of the following subpartitions of seen and unseen features: features seen for the first time at a given time t_j (for $j = 0, \ldots, T$), unseen features alive at time t_0 and unseen features born between any two consecutive times t_j and t_{j+1} (for $j = 0, \ldots, T-1$).

5.2.1 Seen features

As already mentioned, we can apply PG to sample from the posterior trajectories of the seen features. In particular, for features that are seen at time t_0 we can simply apply Algorithm 1 as in the fixed-K approximation by replacing the term $\frac{\alpha\beta}{K}$ with 0. This is possible as observing a feature allocation matrix Z_t updates the prior probability of features as in the posterior beta process (Thibaux and Jordan, 2007), meaning that we can draw each seen feature k from a Beta$(n_{kt},\, \beta + N_t - n_{kt})$ (recall that n_{kt} is the number of objects in which feature k is active at time t).

More generally, consider features that are seen for the first time at a given time t_j. As they cannot be identified individually based on any feature matrix Z_{t_k} for $k < j$, these features need to be drawn from the posterior beta process at time t_j and propagated both forward and backwards. Note that simulating from the W-F diffusion backwards in time is not a problem as each W-F$(0, \beta)$ diffusion is time-reversible with respect to the speed density of the PRF (Griffiths, 2003). The additional backward propagation requires adjusting Algorithm 1, already modified by replacing $\frac{\alpha\beta}{K}$ with 0, by further replacing the steps before the while loop with Algorithm 2, where for simplicity we describe the particular case of features seen for the first time at time t_1. This description can be easily generalized to features that are seen for the first time at a generic time point $t \in \{t_0, \ldots, t_T\}$.

5.2.2 Unseen features

We now describe a thinning scheme to simulate the unseen features alive at time t_0. Denote the slice variable values at each time by $s_{t_0}, \ldots, s_{t_T}$ and note that sampling the set of unseen features from the truncated posterior beta process at time t means drawing samples from a Poisson process on $[s_t, 1)$ with rate measure $x^{-1}(1-x)^{\beta+N_t-1}dx$ (Thibaux and Jordan, 2007), which yields only a finite number of features whose probability is larger than s_t. First, draw the unseen features from the truncated posterior beta process at time t_0. Then, propagate them forward to time t_1 according to the W-F diffusion and accept them with probability $(1-x(t_1))^{N(t_1)}$, namely the binomial likelihood of not seeing them in any object at time t_1. Finally, iterate this propagation and rejection steps up to time t_T. The details of this thinning scheme are given in Algorithm 3.

Notice that simulating the trajectories of the unseen features born between time t_0 and t_1 is equivalent to Algorithm 3 from time t_1 onwards. The only difference is that these features, drawn at time t_1, need to be simulated backwards to time t_0 as well, hence the

Algorithm 2: PG: features seen for the first time at time t_1.

Input: Reference trajectory $x^r_{t_0:t_T}$; M.
Set $x^M_{t_1} = x^r_{t_1}$;
Draw $x^i_{t_1} \sim \text{Beta}(n_{t_1},\, \beta + N_{t_1} - n_{t_1})$ for $i = 1, \ldots, M-1$;
Set $x^M_{t_0} = x^r_{t_0}$;
Simulate $x^i_t | x^i_{t_1} \sim \text{WF}(0, \beta)$ backwards for $t \in [t_1, t_0]$ for $i = 1, \ldots, M-1$;
Set $x^M_{t_2} = x^r_{t_2}$;
Simulate $x^i_t | x^i_{t_1} \sim \text{WF}(0, \beta)$ for $t \in [t_1, t_2]$ for $i = 1, \ldots, M-1$;
Compute $w^i_{t_0} = (1 - x^i_{t_0})^{N_{t_0}}$ for $i = 1, \ldots, M$;
Compute $w^i_{t_2} = (x^i_{t_2})^{n_{t_2}} (1 - x^i_{t_2})^{N_{t_2} - n_{t_2}} w^i_{t_0}$ for $i = 1, \ldots, M$;
Draw $\bar{x}^i_{t_2}$ with $P(\bar{x}^i_{t_2} = x^i_{t_2}) \propto w^i_{t_2}$ for $i = 1, \ldots, M-1$;
Set $\bar{x}^M_{t_2} = x^r_{t_2}$;
Simulate $x^i_t | x^i_{t_2} \sim \text{WF}(0, \beta)$ for $t \in [t_2, t_3]$ for $i = 1, \ldots, M-1$;
Set $j \leftarrow 3$;

Algorithm 3: Thinning: unseen features alive at time t_0

Draw from a Poisson process on $[s_{t_0}, 1)$ with rate measure $\alpha x^{-1}(1-x)^{\beta + N_{t_0} - 1} dx$ and denote by $\{x^i_{t_0}\}_{i \in A}$ the resulting candidate particles;
Set $j \leftarrow 1$;
while $t_j < t_T$ **do**
  Simulate $x^i_t | x^i_{t_j} \sim \text{WF}(0, \beta)$ for $t \in [t_j, t_{j+1}]$ for all $i \in A$;
  Accept $x^i_{t_{j+1}}$ with probability $(1 - x^i_{t_{j+1}})^{N_{t_{j+1}}}$ for all $i \in A$;
  Remove from A the indices of the rejected particles;
  Set $j \leftarrow j + 1$;
end
Output: Trajectories $\{x^i_{t_0:t_T}\}_{i \in A}$ of the unseen features alive at time t_0 from the truncated $\text{PRF}(\alpha, \beta)$.

Algorithm 4: Thinning: unseen features born between time t_0 and t_1

Draw from a Poisson process on $[s_{t_1}, 1)$ with rate measure $\alpha x^{-1}(1-x)^{\beta+N_{t_1}-1}dx$ and denote by $\{x^i_{t_1}\}_{i\in A}$ the resulting candidate particles;
Simulate $x^i_t|x^i_{t_0} \sim \text{WF}(0,\beta)$ for $t \in [t_0, t_1]$ for all $i \in A$;
for *all* $i \in A$ **do**
  if $x^i_{t_0} > s_{t_0}$ **then**
    Reject $x^i_{t_0}$;
    Set $A \leftarrow A \setminus \{i\}$;
  else
    Accept $x^i_{t_0}$ with probability $(1-x^i_{t_0})^{N_{t_0}}$;
  end
end
Set $j \leftarrow 1$;
while $t_j < t_T$ **do**
  Simulate $x^i_t|x^i_{t_j} \sim \text{WF}(0,\beta)$ for $t \in [t_j, t_{j+1}]$ for all $i \in A$;
  Accept $x^i_{t_{j+1}}$ with probability $(1-x^i_{t_{j+1}})^{N_{t_{j+1}}}$ for all $i \in A$;
  Remove from A the indices of the rejected particles;
  Set $j \leftarrow j+1$;
end
Output: Trajectories $\{x^i_{t_0:t_T}\}_{i\in A}$ of the unseen features born between time t_0 and t_1 from the truncated PRF(α,β).

additional backward simulation followed by the rejection step in the for loop of Algorithm 4. If a feature that is simulated backwards from time t_1 to t_0 has probability 0 by time t_0, then it is a newborn feature and is accepted with probability 1. On the other hand, if its probability at time t_0 is between 0 and s_{t_0}, the particle belongs to the category of features that were alive and unseen at time t_0. Accepting them with probability $(1-x(t_0))^{N_{t_0}}$ compensates for the features that were below the truncation level s_{t_0} in Algorithm 3 and were thus not simulated at time t_0. In this way, only the features whose mass is below the slice variables s_t at all times $t \in \{t_0, \dots, t_T\}$ are not simulated. The exactness of the overall MCMC scheme is preserved by the fact that those features are inactive in all the feature allocation matrices by the definition of slice variable.

Finally note that, for simplicity's sake, Algorithm 4 describes only how to simulate the unseen features that were born between times t_0 and t_1, but the procedure needs to be generalized to account for the features born between any two consecutive time points t_j and t_{j+1}, where $j = 0, \dots, T-1$. In order to do this, it is sufficient to draw the candidate particles at every time t_{j+1}, with $j = 0, \dots, T-1$, propagate them backwards until time t_0 and thin them as follows: if their mass exceeds s_t at any $t \in \{t_0, \dots, t_T\}$, then they are rejected; otherwise, at each backward propagation to time $t \in \{t_0, \dots, t_{T-1}\}$ they are accepted with probability $(1-x(t))^{N_t}$.

6. Application: linear-Gaussian likelihood model

The WF-IBP we have described defines a prior over latent feature allocation matrices Z and the corresponding feature probabilities X exhibiting a dependency structure over time. The next step is to relate Z to the observed data by means of a given likelihood model. The first model we explore is the linear-Gaussian likelihood, a very common choice in latent feature models (e.g., Doshi-Velez and Ghahramani, 2009; Griffiths and Ghahramani, 2011; Gershman et al., 2015).

Assume that the collection of observations O_t at time $t = t_0, \ldots, t_T$ is in the form of an $N \times D$ matrix generated by the matrix product $O_t = Z_t \times A + \epsilon_t$. Z_t is the $N \times K$ binary matrix of feature assignments at time t and A is a $K \times D$ factor matrix whose rows represent the feature parameters ρ. The matrix product is the way Z_t determines which features are active in each observation, and ϵ_t is a $N \times D$ Gaussian noise matrix, whose entries are assumed to be distributed as independent $\mathcal{N}(0, \sigma_X^2)$. A typical inference problem is to infer both the feature allocation matrices Z and the factor matrix A. In order to achieve this, we place on each element of A an independent prior $\mathcal{N}(0, \sigma_A^2)$ and on the hyper-parameter σ_A^2 an inverse-gamma prior $\Gamma^{-1}(1, 1)$. This choice of priors is convenient as it is easy to obtain the posterior distributions of σ_A^2 and A (for the case $T = 1$, see Doshi-Velez and Ghahramani, 2009).

For simplicity of notation, consider a fixed number of features K. Denote by $\bar{Z}$ the $TN \times K$ matrix obtained by concatenating the feature matrices Z vertically, and by $\bar{O}$ the $TN \times D$ matrix obtained by combining the observations $\{O_t\}_{t=t_0}^{t_T}$ in the same way. The posterior of A is matrix Gaussian with the following mean μ^A (a $K \times D$ matrix) and, for each column of A, the following covariance matrix Σ^A (a $K \times K$ matrix).

$$\mu^A = \left(\bar{Z}^T \bar{Z} + \frac{\sigma_X^2}{\sigma_A^2} I\right)^{-1} \bar{Z}^T \bar{O}$$
$$\Sigma^A = \sigma_X^2 \left(\bar{Z}^T \bar{Z} + \frac{\sigma_X^2}{\sigma_A^2} I\right)^{-1}.$$

By conjugacy, the posterior distribution for σ_A^2 is still inverse gamma with updated parameters, namely

$$\sigma_A^2 \sim \Gamma^{-1}\left(1 + \frac{1}{2}KD, 1 + \frac{1}{2}\sum_k \sum_d A_{kd}^2\right).$$

6.1 Simulations and results

We tested the WF-IBP combined with a linear-Gaussian likelihood on a variety of synthetic data sets. Starting from the fixed-K approximation, we generated $N = 50$ observations at each of 40 equally-spaced time points as in the linear-Gaussian model. The true factor matrix A contained $K = 3$ latent features in the form of binary vectors of length $D = 30$. Their probability of being active was determined continuously over time by three independent W-F$(1, 1)$ diffusions, simulated for 0.01 diffusion time-units between every two consecutive time points. The resulting observations were corrupted by a large amount of noise ($\sigma_X = 0.5$). 1000 iterations of the overall algorithm were performed, choosing a

burn-in period of 100 iterations and setting the time-units and drift parameters of the W-F diffusion equal to the true ones in the PG update. As ground truth was available, we were able to test the ability of the algorithm to recover the true feature allocation matrices, the latent feature parameters and their probabilities over time.

Figure 4-top-left compares the true underlying feature matrices at times $t = \{1, 14, 27, 40\}$ in terms of the most frequently active features in the posterior mean matrices, where a feature is set to be active if that is the case in more than half of the samples of the Markov chain. The resulting mean matrices almost perfectly match the true underlying feature matrices. Figure 4-top-right compares the trajectories over time of the true feature probabilities with the inferred ones. The latter tend to be less than two standard deviations away from the former, meaning that the true feature trajectories are closely tracked. Figure 4-bottom-left compares the three features represented by the true factor matrix A and the ones in the posterior mean matrix $\hat{A}$, showing that the algorithm was able to recover accurately the hidden features underlying the noisy observations. Figure 4-bottom-right plots the log-likelihood at each iteration, showing that the algorithm converged quickly, namely in fewer than 50 iterations.

Then, we tested the ability of the slice sampler-based algorithm to recover the correct number of latent features when given a similar set of synthetic data, this time consisting of 4 latent features evolving over 6 time points. The algorithm was initialized with one feature and run for 3300 iterations with a burn-in period of 1000 iterations. As in the finite case, the true underlying feature allocation matrices and feature probabilities were closely recovered as illustrated by the top row of Figure 5. The bottom row of Figure 5 shows that the features were reconstructed accurately and their correct number detected in about 700 iterations.

Finally, we focused on the ability of Particle Gibbs to track the feature trajectories in a wider variety of settings and under model misspecification. We generated a set of feature allocation matrices obtained by letting three feature probabilities evolve over time, first by simulating standard W-F diffusions and then by introducing jumps and spikes. Figure 6 confirms the robustness of the algorithm in the presence of mismatches between the W-F diffusion and the true process determining the feature trajectories. We simulated a set of 20 feature allocation matrices under an increasing amount of mismatch, first by introducing a jump of increasing size from time 10 to 11, and then by adding a spike of increasing sharpness at time 10. Figure 6-left shows that, although larger jump sizes lead to larger discrepancies between the true and inferred trajectories around the jump, the former are always less than two standard deviations away from the latter. Figure 6-right shows that in all three cases the algorithm is able to detect the presence of the spike at time 10. Figure 7 illustrates the performance of the algorithm on a decreasing number of observations. The results show that the posterior mean of the target distribution closely corresponds to the true feature probabilities and, given a sufficient number of observations, always fall within the interval given by two standard deviations about the posterior mean.

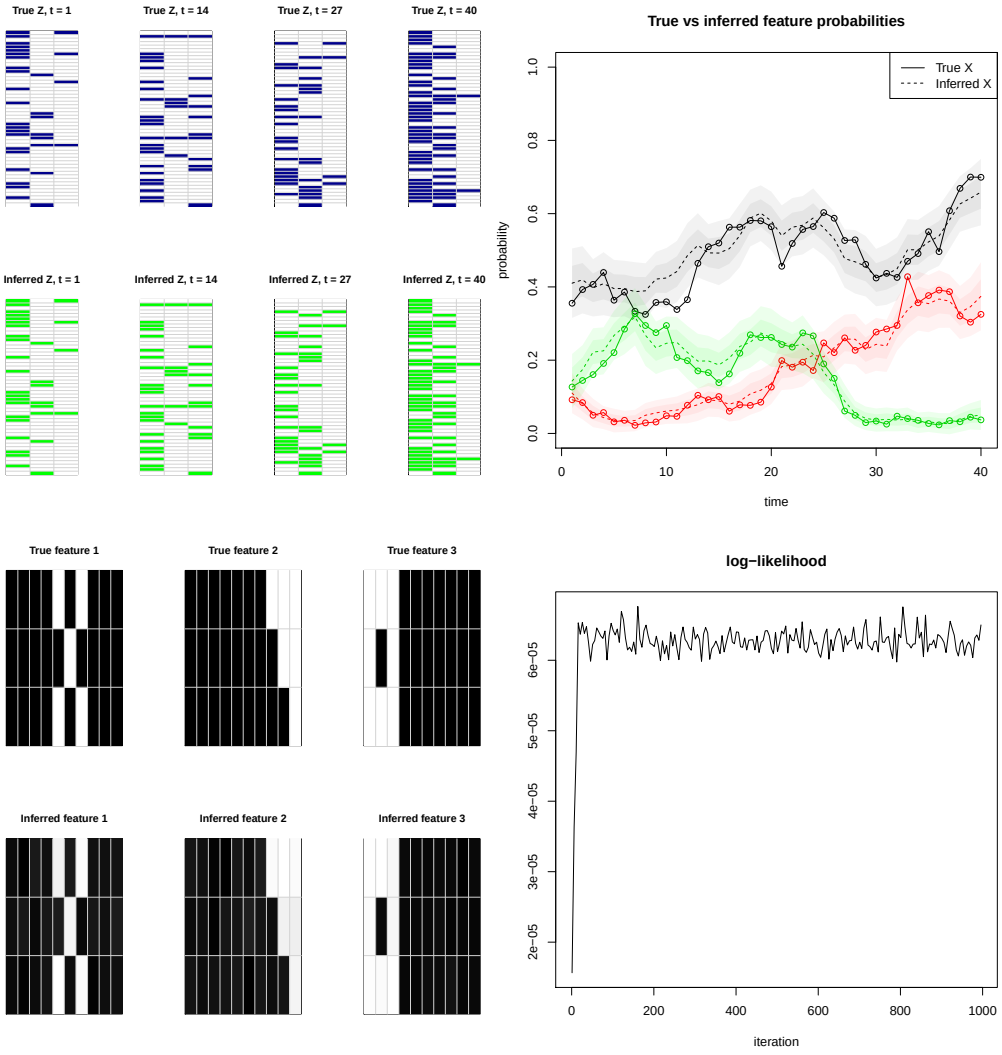


Figure 4: Fixed-K approximation. Top-Left: Subset of true feature allocation matrices vs inferred ones. Top-Right: True vs inferred feature probabilities over time (the dark and the light shaded areas respectively indicate one and two standard deviations about the posterior mean). Bottom-Left: True vs inferred features (black and white entries respectively correspond to 0 and 1, while the shades of grey to the values in between). Bottom-Right: Convergence of the log-likelihood of the observed data.

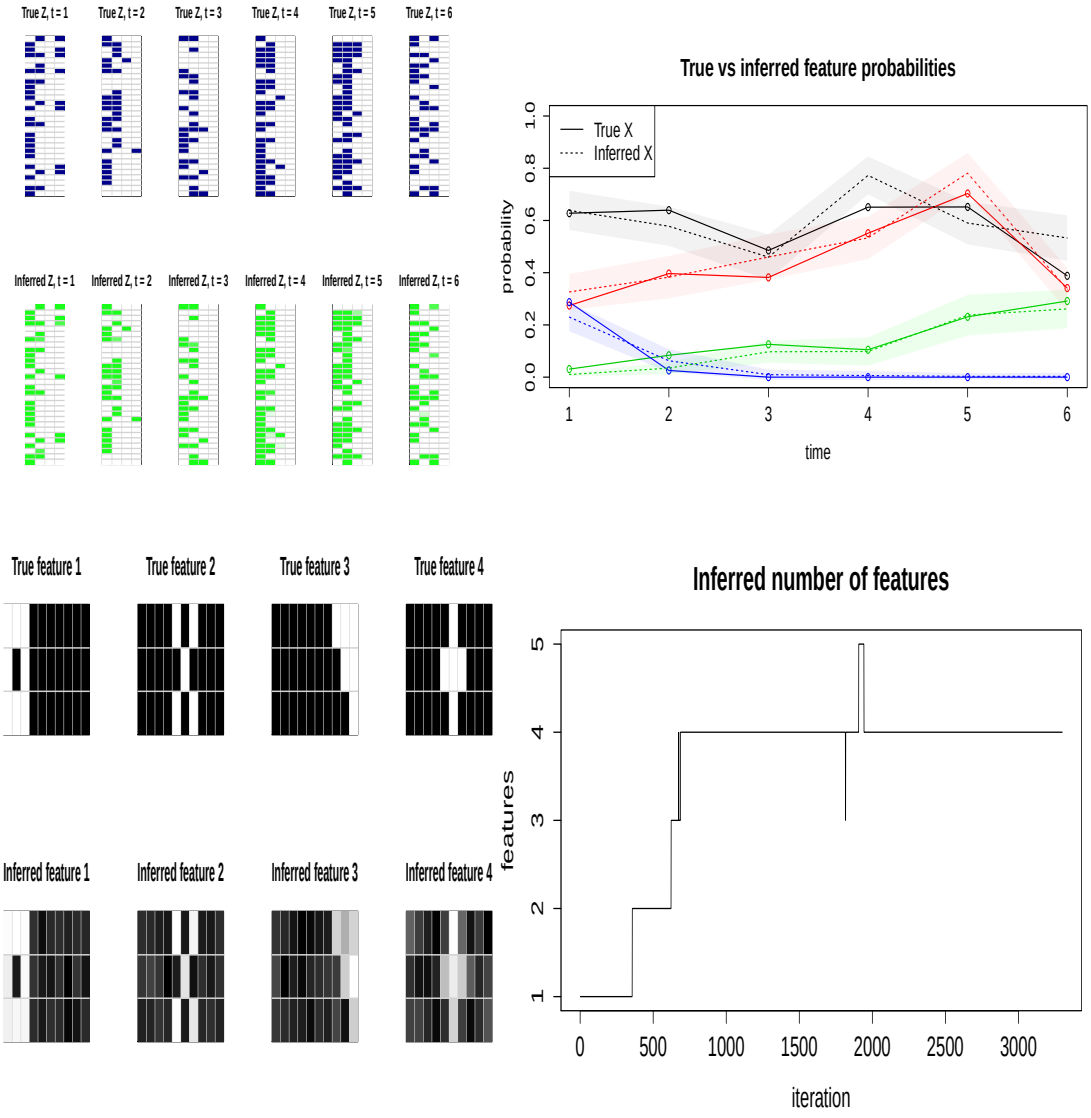


Figure 5: WF-IBP. Top-Left: True vs inferred feature allocation matrices. Top-Right: True vs posterior mean feature trajectories (the shaded areas represent one standard deviation). Bottom-Left: Comparison between true and inferred features. Bottom-Right: Convergence to the true number of features.

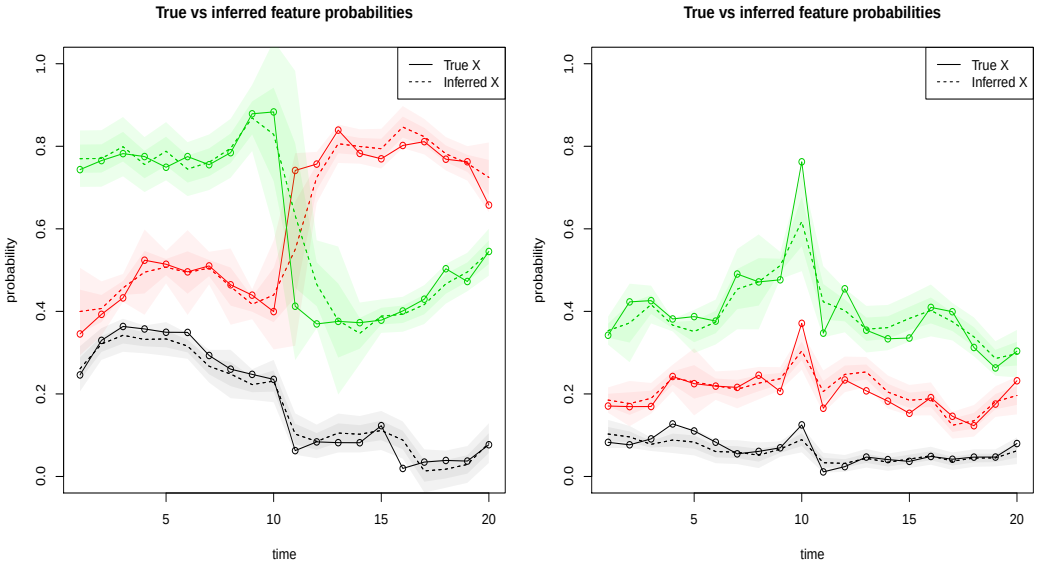


Figure 6: Comparison between the posterior means and the true feature probabilities under an increasing amount of model misspecification. The dark and light shaded areas respectively correspond to 1 and 2 standard deviations about the posterior mean. Left: Jump of increasing size between times 10 and 11. Right: Spike of increasing sharpness at time 10.

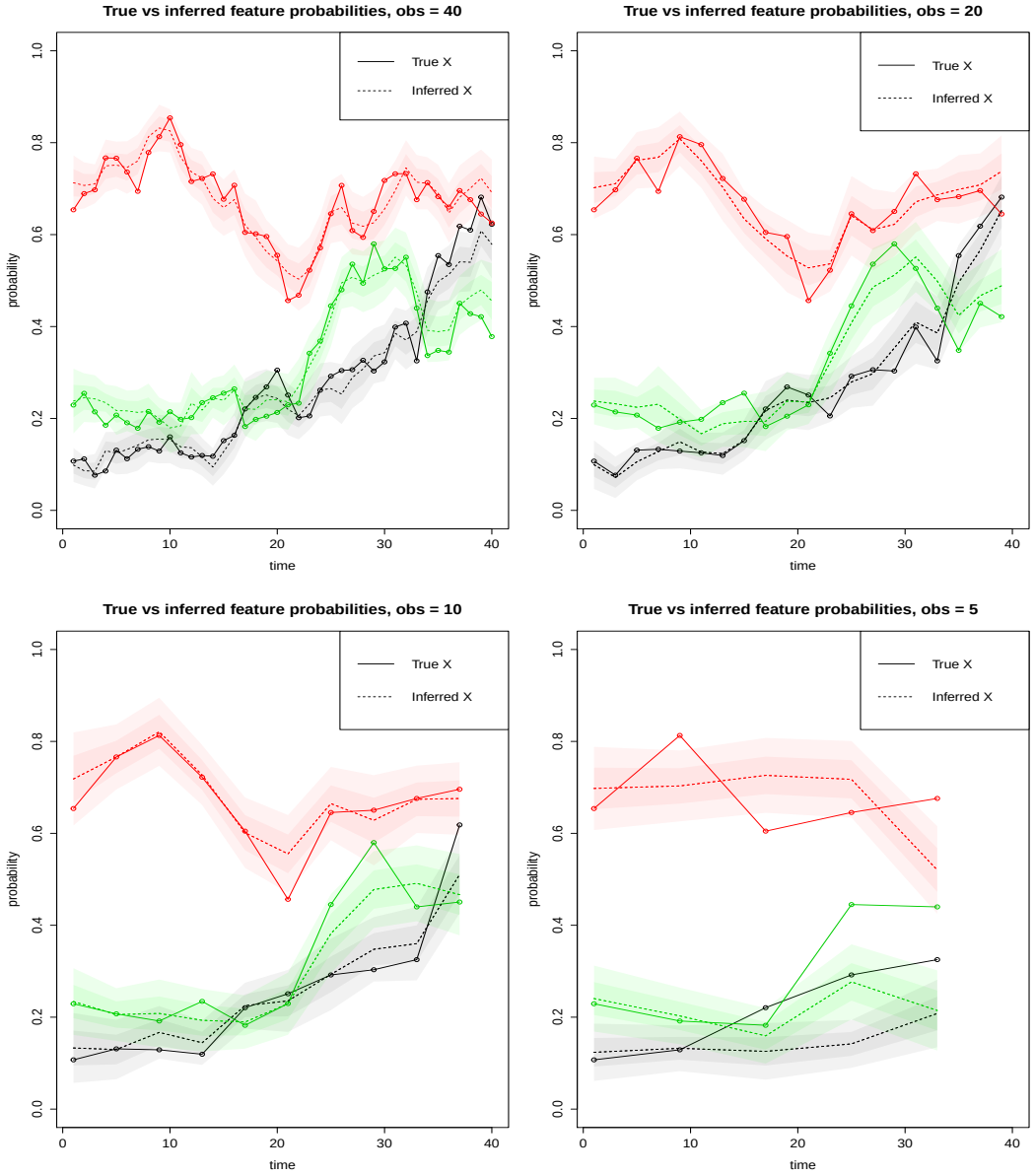


Figure 7: Comparison between true feature probabilities and posterior means obtained via Particle Gibbs for a varying number of observations. The dark and light shaded areas respectively correspond to 1 and 2 standard deviations about the posterior mean.

7. Topic modeling application

In this section we apply the WF-IBP to the modeling of corpora of time-stamped text documents. This is a natural application as documents can be seen as arising from an unknown number of latent topics whose popularity is evolving over time. A related model to achieve this goal is described in Blei and Lafferty (2006), where a Gaussian state space model captures the evolution of topics in such a way that both the content of topics and their proportions evolve over time. This work has had a great impact in the topic modeling community, and anticipates a number of directions for future work that we address here. Specifically their model is parametric, in that the number of topics K is fixed and needs to be pre-specified. The authors claim that it would desirable to drop this assumption to have a more flexible model and, in particular, foresee a process involving births and deaths of topics. The WF-IBP topic model elegantly achieves these goals. Unlike Blei and Lafferty (2006) we focus on the evolution of topic probabilities rather than topic contents, noting that the modeling of time-varying topic contents is orthogonal to our work and could be incorporated into the WF-IBP in future developments. Another class of models called Dirichlet processes aim at modeling the evolution of topics in a time-dependent and nonparametric way. Some of these models, however, assume the evolution of topic probabilities to be unimodal (e.g., Rao and Teh, 2009), while others are HDP-based (Ahmed and Xing, 2012; Dubey et al., 2013) and implicitly assume a positive correlation between the probability of a topic being active and the proportion of that topic within each document. Coupling topic proportions and topic probabilities is undesirable as rare topics may account for a large proportion of words in the few documents in which they appear. Our nonparametric topic model decouples the probability of a topic and its proportion within documents and offers a flexible way to model topic evolutions over time. We achieve this by incorporating time-dependency into the focused topic model presented in Williamson et al. (2010b), which makes use of the IBP to select the finite number of topics that each document treats.

7.1 The WF-IBP topic model

First consider the case in which the number of topics K underlying the corpus of seen documents is known. Define topics as probability distributions over a dictionary of D words and model them as $(\rho_k)_{k=1}^K \overset{iid}{\sim} \text{Dirichlet}(\bar{\eta})$, given a vector $\bar{\eta}$ of length D. Let ρ be the resulting vector and assume the components of $\bar{\eta}$ to be all equal to a constant $\eta > 0$. Consider the usual setting in which the time-dependent popularity of topic (feature) k is denoted by X_k and the binary variables Z_{ikt} indicate whether document i contains topic k at time t. Then, for all $t = t_0, \ldots, t_T$ and $k = 1, \ldots, K$, sample

$$\begin{aligned}
\theta_{it} \mid \{Z_{it} = z_{it}, \phi_t = \phi_t'\} &\sim \text{Dirichlet}(z_{it} \circ \phi_t'), \quad \forall i = 1, \ldots, N_t, \\
\phi_{kt} &\sim \text{Gamma}(\gamma, 1), \\
(Z_{ikt})_{i=1}^{N_t} \mid \{X_k = x_k(t)\} &\overset{iid}{\sim} \text{Bernoulli}(x_k(t)), \\
X_k &\sim \text{WF}\left(\frac{\alpha\beta}{K}, \beta\right),
\end{aligned}$$

where ϕ_{kt} is the kth component of ϕ_t, a K-long vector of topic proportions, and θ_{it} the ith row of θ_t, a $N_t \times K$ matrix with the distributions over topics for each document at

time t. The operation $z_{it} \circ \phi_t$ stands for the Hadamard product between z_{it} and ϕ_t and the Dirichlet is defined over the positive components of the resulting vector. While the topic allocation matrix Z_t encodes which subset of the K topics appears in each document at time t, the variables ϕ_{kt} are related to the proportion of words that topic k explains within each document. Unlike HDP-based models, these two quantities are here modeled independently.

For every document $i = 1, \ldots, N_t$, draw the total number of words from a negative-binomial $W_{it} \sim \mathrm{NB}(\sum_k z_{ikt}\phi_{kt}, 1/2)$ and, for each word w_{ilt}, $l = 1, \ldots, W_{it}$, sample first the topic assignment

$$a_{ilt} \mid \{\theta_{it} = \theta'_{it}\} \sim \text{Categorical}(\theta'_{it})$$

and then the word

$$w_{ilt} \mid \{a_{ilt} = a'_{ilt}, \rho = \rho'\} \sim \text{Categorical}(\rho'_{a'_{ilt}}).$$

Assume now that the number of potential topics K needs to be learned from the data. The nonparametric extension of this model is easily obtained by replacing the process generating the topic allocation matrices with the WF-IBP, so that topics arise as in the PRF and evolve as independent $\mathrm{WF}(0, \beta)$. The feature allocation matrices can be drawn as described in Section 8. In this way, we obtain a time-dependent extension of the IBP compound Dirichlet process presented in Williamson et al. (2010b).

7.2 Posterior inference

In order to infer the latent variables of the model, it is convenient to integrate out the parameters ρ and θ. This can be done easily thanks to the conjugacy between the Dirichlet and the Categorical distribution. In this way, we can run a Gibbs sampler for posterior inference only over the remaining latent variables and are able to follow the derivation of conditionals given by Williamson et al. (2010a). Note that, in our case, we have introduced the slice variable and do not integrate out the topic allocation matrix. Denote by W the complete set of words and by A the complete set of topic assignments a_{ilt} for all times $t = t_0, \ldots, t_T$, documents $i = 1, \ldots, N_t$ and words $l = 1, \ldots, W_{it}$. Denote by S_t the slice variable and by W_t the complete set of words at time t. The conditional distributions that we need to sample from for all times $t = t_0, \ldots, t_T$ are

$$\begin{aligned}
&p(A_t \mid Z_t, w, \phi_t),\\
&p(\phi_t, \gamma \mid A_t, W_t, Z_t),\\
&p(S_t \mid Z_t, X(t)),\\
&p(Z_t \mid A_t, X(t), \phi_t, S_t).
\end{aligned}$$

Conditioning on all the other topic assignments a_{-il}, each topic assignment a_{ilt} can be sampled from

$$p(a_{ilt} = k \mid a_{-il}, Z_t, W, \phi_t) \propto (n_k^{w_{il}} + \eta)\frac{n_{kt}^i + \phi_{kt} z_{ikt}}{n_k + \eta D - 1},$$

where $n_k^{w_{il}}$ denotes the number of times that word w_{il} has been assigned to topic k excluding assignment a_{il}, n_{kt}^i the number of words assigned to topic k in document i excluding assignment a_{il} and n_k the total number of words assigned to topic k.

After placing a hyper-prior on $p(\gamma)$, we can sample ϕ and γ via a Metropolis-Hastings step. Indeed, we know that

$$p(\phi_{kt}, \gamma \mid A, Z_t) \propto \frac{\phi_{kt}^{\gamma-1} e^{-\phi_{kt}}}{\Gamma(\gamma)} p(\gamma) \prod_{i=1}^{N_t} \frac{\Gamma(\phi_{kt} z_{ikt} + n_{kt}^i)}{\Gamma(\phi_{kt} z_{ikt}) n_{kt}^i! 2^{\phi_{kt} z_{ikt} + n_{kt}^i}}.$$

Conditioning on Z_t, the slice variable is sampled according to its definition:

$$p(S_t \mid Z_t, X(t)) = \frac{1}{x^{min}(t)} 1_{s_t}([0, x^{min}(t)]),$$

where $x^{min}(t)$ is the minimum among the probabilities of the active topics at time t. As for the feature allocation matrices Z, we sample only the finite number of its components whose topic probability $x_k(t)$ is greater than the slice variable S_t. Assume we are sampling each entry Z_{ikt} sequentially and denote respectively by $x_1^{min}(t)$ and $x_0^{min}(t)$ the minimum active topic probability in the cases $Z_{ikt} = 1$ and $Z_{ikt} = 0$. Let n_{ikt} denote the total number of words assigned to topic k in document i at time t. Then we have that

$$p(Z_{ikt} = 1 \mid A, x_k(t), \phi_{kt}) = \begin{cases} 1, & \text{if } n_{ikt} > 0 \\ \frac{x_k(t) x_0^{min}(t)}{x_k(t) x_0^{min}(t) + 2^{\phi_{kt}} (1 - x_k(t)) x_1^{min}(t)}, & \text{if } n_{ikt} = 0. \end{cases}$$

$$p(Z_{ikt} = 0 \mid A, x_k(t), \phi_{kt}) = \begin{cases} 0, & \text{if } n_{ikt} > 0 \\ \frac{2^{\phi_{kt}} (1 - x_k(t)) x_1^{min}(t)}{x_k(t) x_0^{min}(t) + 2^{\phi_{kt}} (1 - x_k(t)) x_1^{min}(t)}, & \text{if } n_{ikt} = 0. \end{cases}$$

The full conditional distributions presented so far are derived in Appendix C. As in the linear-Gaussian case, when considering the probability of setting $Z_{ikt} = 1$ for a feature that is currently inactive across all observations at time t, it is necessary to jointly propose a new value ϕ_{kt} by drawing it from its prior distribution Gamma$(\gamma, 1)$. Finally, inference on the trajectories of the topic probabilities is a direct application of the Particle Gibbs and thinning scheme outlined for the general case.

7.3 Topic model: simulation and results

At each of 4 time points, a small corpus of $N = 30$ documents was simulated by selecting up to $K = 4$ latent topics for each document and picking words from a dictionary of $D = 100$ words. The hyper-parameter of the Dirichlet prior over words was chosen to be a vector with components equal to $\eta = 0.1$, a Gamma(5,1) hyper-prior was placed on γ and we let topic probabilities evolve as independent W-F$(1, 1)$ diffusions with 0.1 diffusion time-units between each observation. Assume K is known and focus on inference over the remaining parameters. Fix the time-units and drift parameters of the W-F diffusion to their true values in the PG update. We ran the Gibbs sampler for 3000 iterations with a burn-in period of 300 iterations. The log-likelihood converged in about 500 iterations (Figure 8)

and the algorithm was able to infer closely the latent topic allocation matrices (Figure 9-left). The percentage of words assigned to the correct topics were 81% at t_1, 82% at t_2, 83% at t_3 and 85% at t_4.

A Monte Carlo estimate for the probability of word $w = 1, \ldots, D$ under topic k is

$$\hat{\rho}_{kw} = \frac{n_k^w + \eta}{n_k + D\eta},$$

where n_k^w is the number of times word w has been assigned to topic k, n_k is the total number of words assigned to topic k and D is the number of words in the dictionary. A Monte Carlo estimate for the probability of topic k in document i is

$$\hat{\theta}_{ikt} = \frac{n_{ikt} + z_{ikt}\phi_{kt}}{\sum_k (n_{ikt} + z_{ikt}\phi_{kt})}. \tag{11}$$

These two quantities are given by the posterior mean of the Dirichlet distribution under a categorical likelihood. $\hat{\rho}_{kw}$ has been used to plot the posterior distribution over words in Figure 9-right. These results confirm the ability of the algorithm to recover ground truth and provide useful information both at word and topic level.

Finally, we tested the ability of the algorithm to reconstruct topics when presented with a decreasing number of observed documents. 100 documents were simulated at each of two time points as described above. Figure 10 shows how well the probability of the 10 most likely words of the first topic was reconstructed for $N = 60$, 40, 20 and 10 observed documents per time point. As expected, the more documents are observed the more accurate the reconstruction of topics is, with a drop in performance when only 10 documents per time point are observed.

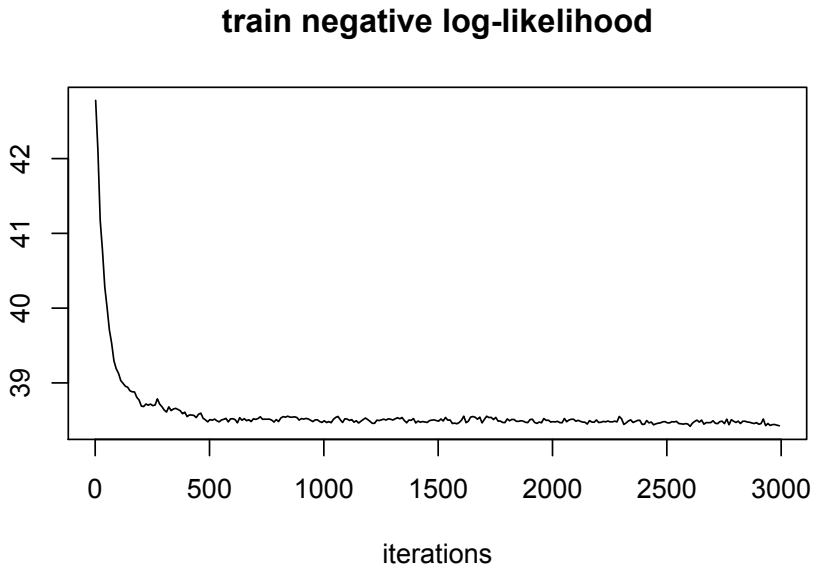


Figure 8: Convergence of the train negative log-likelihood.

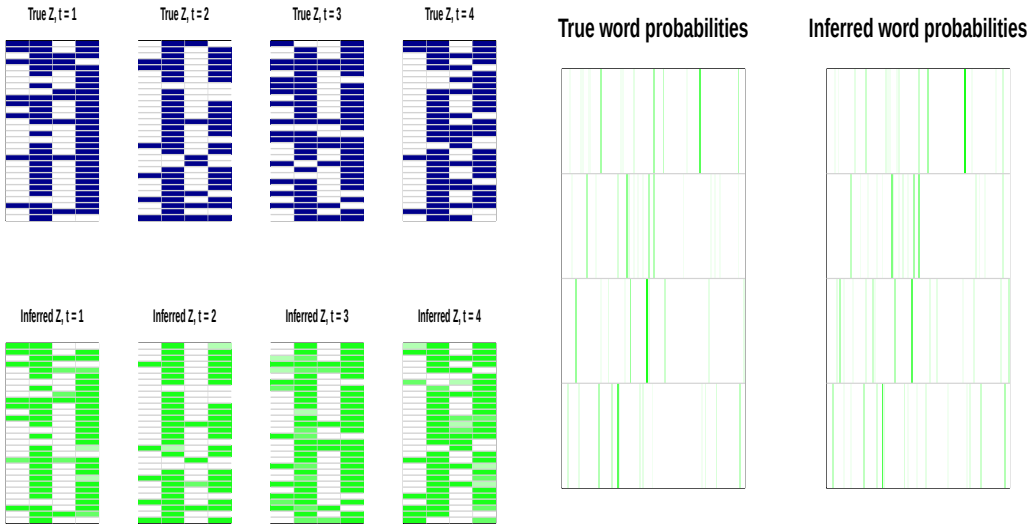


Figure 9: Left: Comparison between the true and the posterior mean topic allocation matrices at each time. Right: True vs inferred distributions over words for each topic. Each row is a topic ($K = 4$) and each column is a word from the dictionary ($D = 100$) (the darker the green, the larger the probability of the corresponding word).

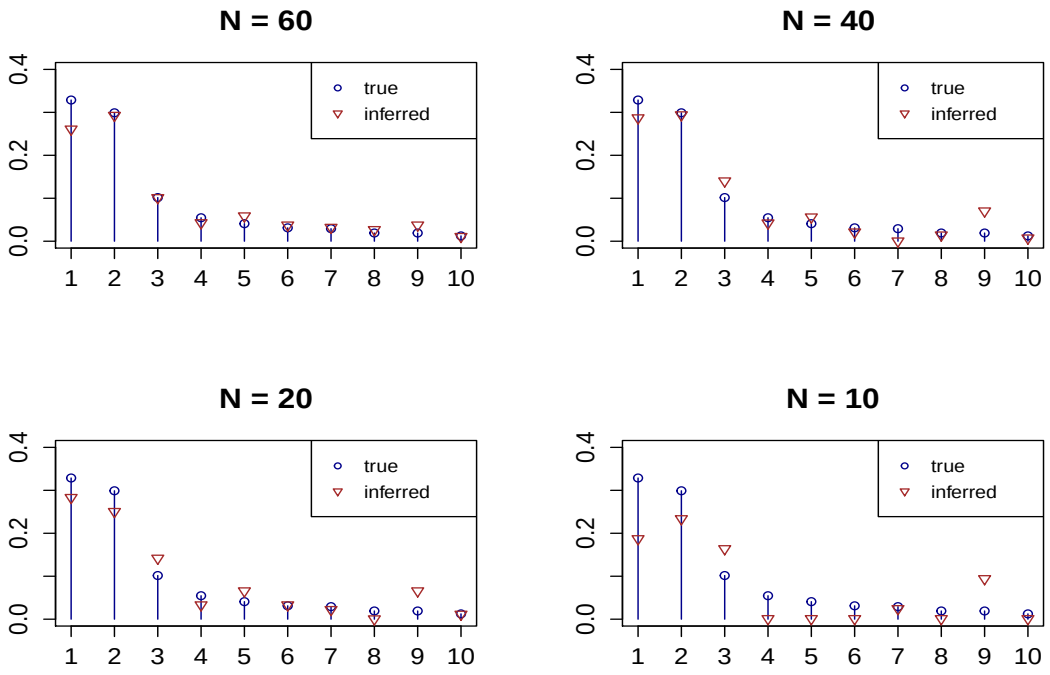


Figure 10: Comparison, given different numbers N of observed documents, between the true probabilities of the 10 most likely words within a given topic and the inferred probabilities of those words for that topic. The more documents are observed, the better the reconstruction of the topic is.

7.3.1 Comparison: static model and hierarchical model

We now demonstrate the advantages of modeling time-dependency by comparing our model with two alternative versions. First, we consider a static counterpart where time is not modeled and thus information about the time-stamp of documents is not exploited. Second, we consider a hierarchical model where feature probabilities at each time point are distributed as conditionally independent beta processes given a lower-layer beta process (Thibaux and Jordan, 2007); note that in this model observations at different time points are modeled separately, but information on the order of the time points is ignored.

In particular, we investigate whether incorporating time into the model improves upon test-set perplexity, a measure widely used in topic modeling settings that assesses the ability of topic models to generalize to unseen data. Given the model parameters Φ, perplexity on documents $D_{test} := \{d_i\}_{i=1}^{M}$ is defined as

$$\text{perplexity}(D_{test} \mid \Phi) = \exp\left(-\frac{\sum_{i=1}^{M} \log p(d_i \mid \Phi)}{\sum_{i=1}^{M} W_i}\right),$$

where W_i denotes the number of words in document d_i. As we assume that words within each document are drawn independently given the model parameters Φ, the probability of document d_i can be computed as

$$p(d_i \mid \Phi) = \prod_{l=1}^{W_i} p(w_{il} \mid \Phi),$$

where

$$p(w_{il} \mid \Phi) = \sum_{k=1}^{K} \theta_{ik}\rho_{kl},$$

recalling that θ_{ik} is the probability of a generic word belonging to topic k in document i and ρ_{kl} is the probability of word l under topic k. These two quantities can be approximated at each iteration of the MCMC algorithm by their current values $\hat{\theta}_{ik}^{(s)}$ and $\hat{\rho}_{kl}^{(s)}$, so that we can approximate the probability of each word by averaging over S samples of the Markov Chain.

$$\hat{p}(w_{il} \mid \Phi) = \frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K} \hat{\theta}_{ik}^{(s)}\hat{\rho}_{kl}^{(s)}.$$

Note that the perplexity is inversely proportional to the likelihood of the data and thus lower values indicate better performance. Chance performance, namely assuming each word to be picked uniformly at random from the dictionary, yields a perplexity equal to the size D of the dictionary.

Different percentages of words were held-out and the model was trained on the remaining data. Testing the model on held-out words is a way to avoid comparing different hyper-parameters, as different treatments of the hyper-parameters could strongly affect the results (Asuncion et al., 2009). For all three models, a dictionary of $D = 1000$ words was used

to generate 30 documents at each of 9 time points. The number of features was fixed to 4 and the algorithms were run 3000 iterations with a burn-in period of 300 iterations. Even though all three algorithms approximately recover the true topic allocations matrices, incorporating time leads to a closer match. This can be measured, for instance, by the Frobenius norm of the difference between each true and inferred Z_t. Table 1 shows that at each time the dynamic model leads to a lower discrepancy with the true topic allocation matrix. Figure 11 shows the posterior trajectories of topic probabilities inferred by the dynamic model and compares them with the posterior feature probabilities inferred by the hierarchical model and the constant values inferred by the static model. It can be observed that the trajectories inferred by the dynamic model are both smoother and closer to the ground truth. Finally, Figure 12 compares the test-set perplexity of the three models (recall that in this case chance performance results in a perplexity of $D = 1000$). As expected, although the hierarchical version performs better than the static counterpart that neglects time-stamp information, it is outperformed by our dynamic model. Indeed, while in the WF-IBP observations that are closer in time exhibit a stronger dependency, the HBP does not explicitly impose an ordering of the time points. The results show that having a suitable model for time dependencies improves on the ability to recover ground truth as well as to generalize to unseen data.

Table 1: Frobenius norm of the difference with the true Z_t. The lower, the better.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
Dynamic	**1.78**	**1.37**	**0.72**	**2.99**	**3.24**	**2.01**	**2.59**	**2.45**	**2.88**
Hierarchical	1.94	1.51	2.80	3.11	3.52	2.91	2.65	2.59	3.24
Static	3.63	4.23	2.84	3.01	3.55	5.27	4.89	3.31	5.90

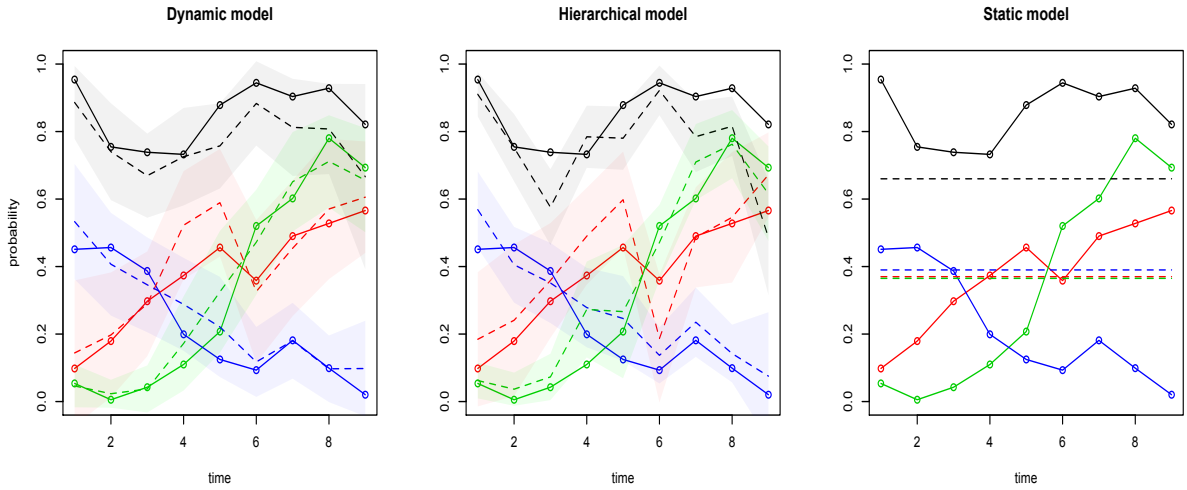


Figure 11: Comparison between true and inferred feature probabilities (respectively continuous and dotted lines) in our fixed-K topic model (left), in the hierarchical version (center) and in the static version (right).

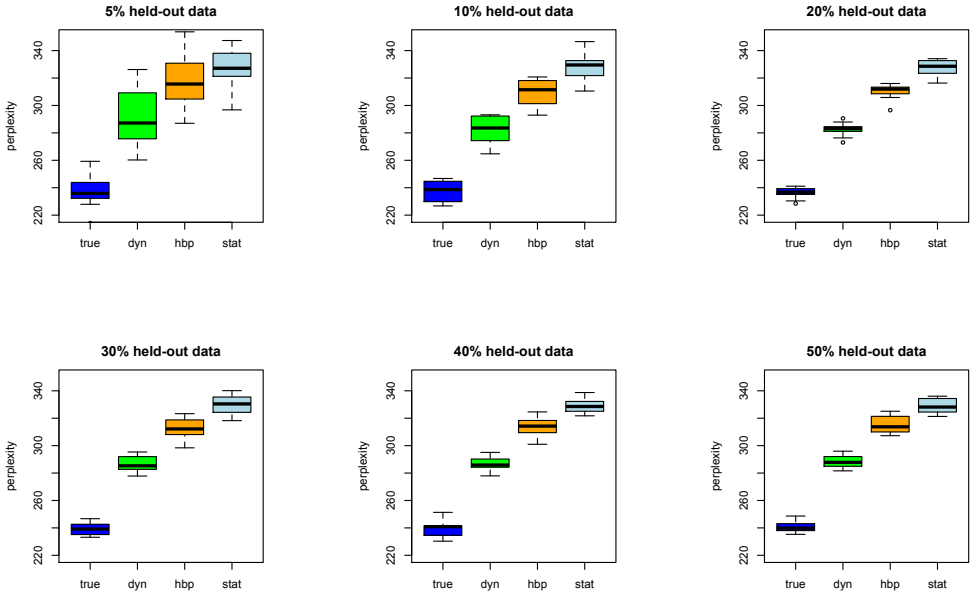


Figure 12: Boxplots of test-set perplexity for different percentages of held-out data for the true model (true), our dynamic model (dyn), the hierarchical version (hbp) and the static version (stat). Each boxplot was obtained by computing the perplexity after holding-out 10 different random subsets of words in the data. Lower values indicate better performance.

7.3.2 Real-world data experiments

We used the WF-IBP topic model to explore the data set consisting of the full text of 5811 NIPS conference papers published between 1987 to 2015.[1] We pre-processed the data and removed words appearing more than 5000 times or fewer than 250 times. The remaining number of word tokens was 4 728 892 with a vocabulary size of 348 672 unique words. Our goal was to discover what topics appear in the corpus and to track the evolution of their popularity over these 29 years.

We set the hyperparameters α and β equal to 1 and the time step to 0.12 diffusion time-units per year so as to reflect realistic evolutions of topic popularity. The Markov chain was run for 2000 iterations with a burn-in period of 200 iterations, setting $\eta = 0.001$ and placing a Gamma(5,1) hyper-prior on γ.

Qualitative results One of the qualitative advantages of modeling time dependency explicitly is that interesting insights into the evolution of topics underlying large collections of documents can be obtained automatically, and uncertainty in the predictions naturally incorporated. The 12 most likely words of 32 topics found in the corpus together with the evolution of their topic proportions are given in Figure 13, where the shaded areas

1. The data set is available at `https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015`.

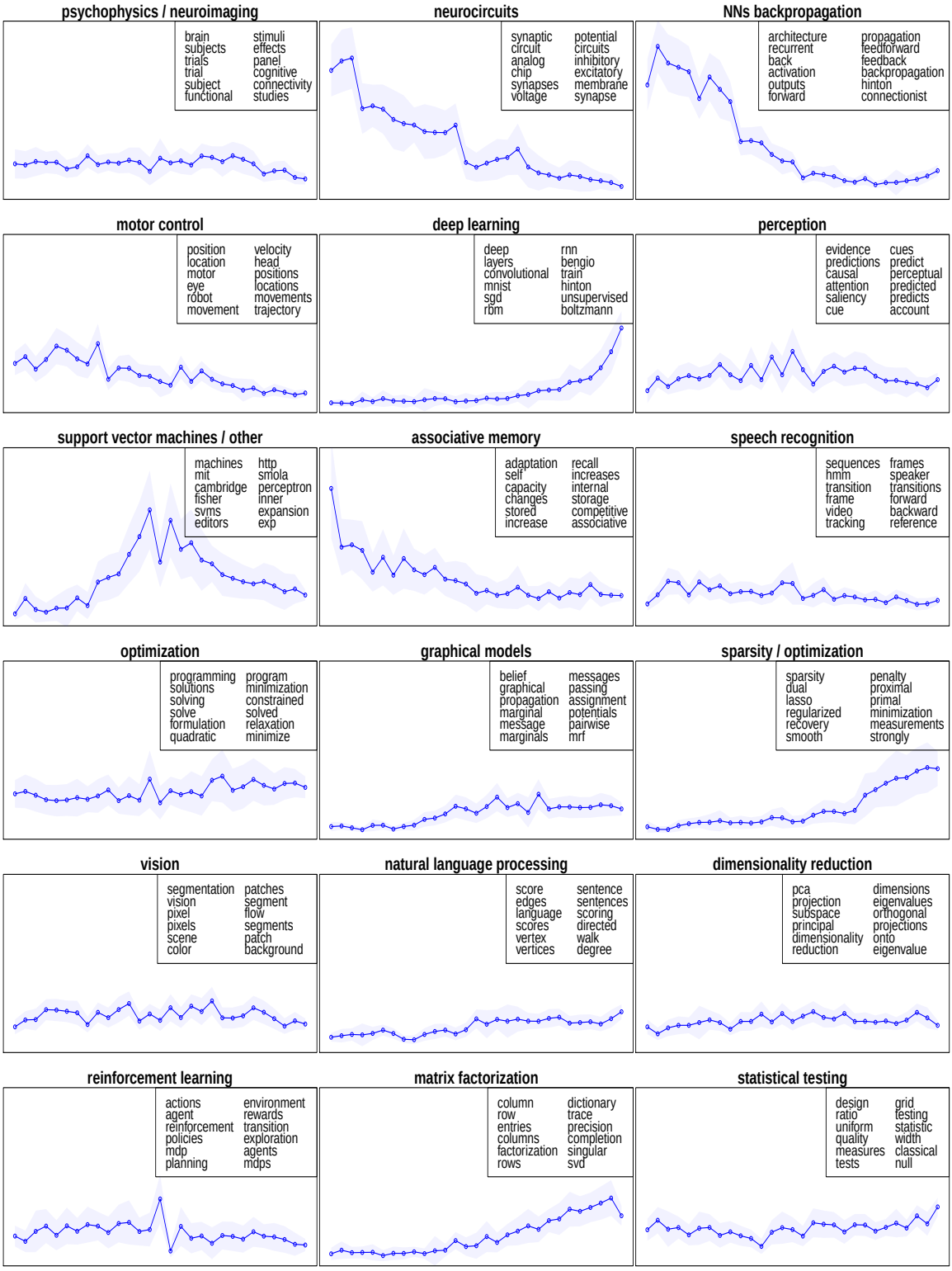
represent one standard deviation around the posterior means. As topics are defined by their distribution over words, it is possible to label them by looking at their most likely words. We observe that, with very few exceptions, the topics detected in the corpus are meaningful and easily interpretable. Figure 14 compares how the popularity of three different approaches to machine learning evolved over time. The results indicate that standard neural networks ('NNs backpropagation') were extremely popular until the early 90s. After this, they went through a steady decline, only to increase in popularity later on. This confirms the well known fact that NNs were largely forsaken in the machine learning community in the late 90s (see LeCun et al., 2015). On the other hand, it can be observed that the popularity of deep architectures and convolutional neural networks ('deep learning') steadily increased over these 29 years, to the point that deep learning became the most popular among all topics in NIPS 2015.

Another key benefit of the WF-IBP over alternative nonparametric topic models is that the overall probability of topics and their proportion within documents are modeled separately, which allows rare topics to be the predominant subject within a few documents. This can be observed by comparing the rarest topic probabilities with the corresponding within-document topic proportions. In a number of documents WF-IBP reveals that the predominant topic is among the rarest topics in the corresponding year, such as in "Text Classification using String Kernels" (information retrieval),"Playing is Believing: The Role of Beliefs in Multi-Agent Learning" (game theory), and "Relative Density Nets: A New Way to Combine Backpropagation with HMMs" (speech recognition).

We then compared the document representations learned by WF-IBP with the ones obtained by Dynamic Topic Models (DTM) (Blei and Lafferty, 2006). One of the benefits of WF-IBP is that it provides sparse and less noisy representations. This is due to the fact that, while DTM assigns a positive probability to all topics within each document, the WF-IBP selects only a subset of topics with positive probability via the feature allocation matrices Z. Recall that DTM requires fixing the number of topics K a priori, hence in our experiments we set $K = 50$. For instance, "Recursive Training of 2D-3D Convolutional Networks for Neuronal Boundary Prediction", "Exploring Models and Data for Image Question Answering" and "Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering" are respectively assigned to deep learning, image recognition and NLP by both models; however, they are respectively explained by 9,8 and 10 topics in WF-IBP, as opposed to 50 in DTM. While it is possible to order the topic proportions in DTM and only consider the ones greater than an arbitrary threshold, WF-IBP automatically sets the probability of irrelevant topics to 0 and offers a more interpretable representation.

Quantitative results We then compared the predictive performance of our fixed-K approximation with its static and hierarchical counterparts. Recall that time-stamps are not used in the static model and their ordering is neglected in the hierarchical model. The results in Figure 15 were obtained by holding out different percentages of words (50%, 60%, 70% and 80%) from all the papers published in 1999 and by training the model over the papers published in the time range 1987-1999. The goal was to investigate whether incorporating time dependency improves the predictions on future documents at time $t+1$ when given the documents up to time t. The held-out words were then used to compute the test-set perplexity after 5 repeated runs with random initializations (the error bars represent

one standard deviation). The dynamic model led to consistently better results, especially as the number of held-out words was increased. This follows from the fact that, the less training data is available in the year in which the models are tested, the more important capturing time-dependence to yield sensible predictions.


psychophysics / neuroimaging
brain stimuli
subjects effects
trials panel
trial cognitive
subject connectivity
functional studies
neurocircuits
synaptic potential
circuit circuits
analog inhibitory
chip excitatory
synapses membrane
voltage synapse
NNs backpropagation
architecture propagation
recurrent feedforward
back feedback
activation backpropagation
outputs hinton
forward connectionist
motor control
position velocity
location head
motor positions
eye locations
robot movements
movement trajectory
deep learning
deep rnn
layers bengio
convolutional train
mnist hinton
sgd unsupervised
rbm boltzmann
perception
evidence cues
predictions predict
causal perceptual
attention predicted
saliency predicts
cue account
support vector machines / other
machines http
mit smola
cambridge perceptron
fisher inner
svms expansion
editors exp
associative memory
adaptation recall
self increases
capacity internal
changes storage
stored competitive
increase associative
speech recognition
sequences frames
hmm speaker
transition transitions
frame forward
video backward
tracking reference
optimization
programming program
solutions minimization
solving constrained
solve solved
formulation relaxation
quadratic minimize
graphical models
belief messages
graphical passing
propagation assignment
marginal potentials
message pairwise
marginals mrf
sparsity / optimization
sparsity penalty
dual proximal
lasso primal
regularized minimization
recovery measurements
smooth strongly
vision
segmentation patches
vision segment
pixel flow
pixels segments
scene patch
color background
natural language processing
score sentence
edges sentences
language scoring
scores directed
vertex walk
vertices degree
dimensionality reduction
pca dimensions
projection eigenvalues
subspace orthogonal
principal projections
dimensionality onto
reduction eigenvalue
reinforcement learning
actions environment
agent rewards
reinforcement transition
policies exploration
mdp agents
planning mdps
matrix factorization
column dictionary
row trace
entries precision
columns completion
factorization singular
rows svd
statistical testing
design grid
ratio testing
uniform statistic
quality width
measures classical
tests null

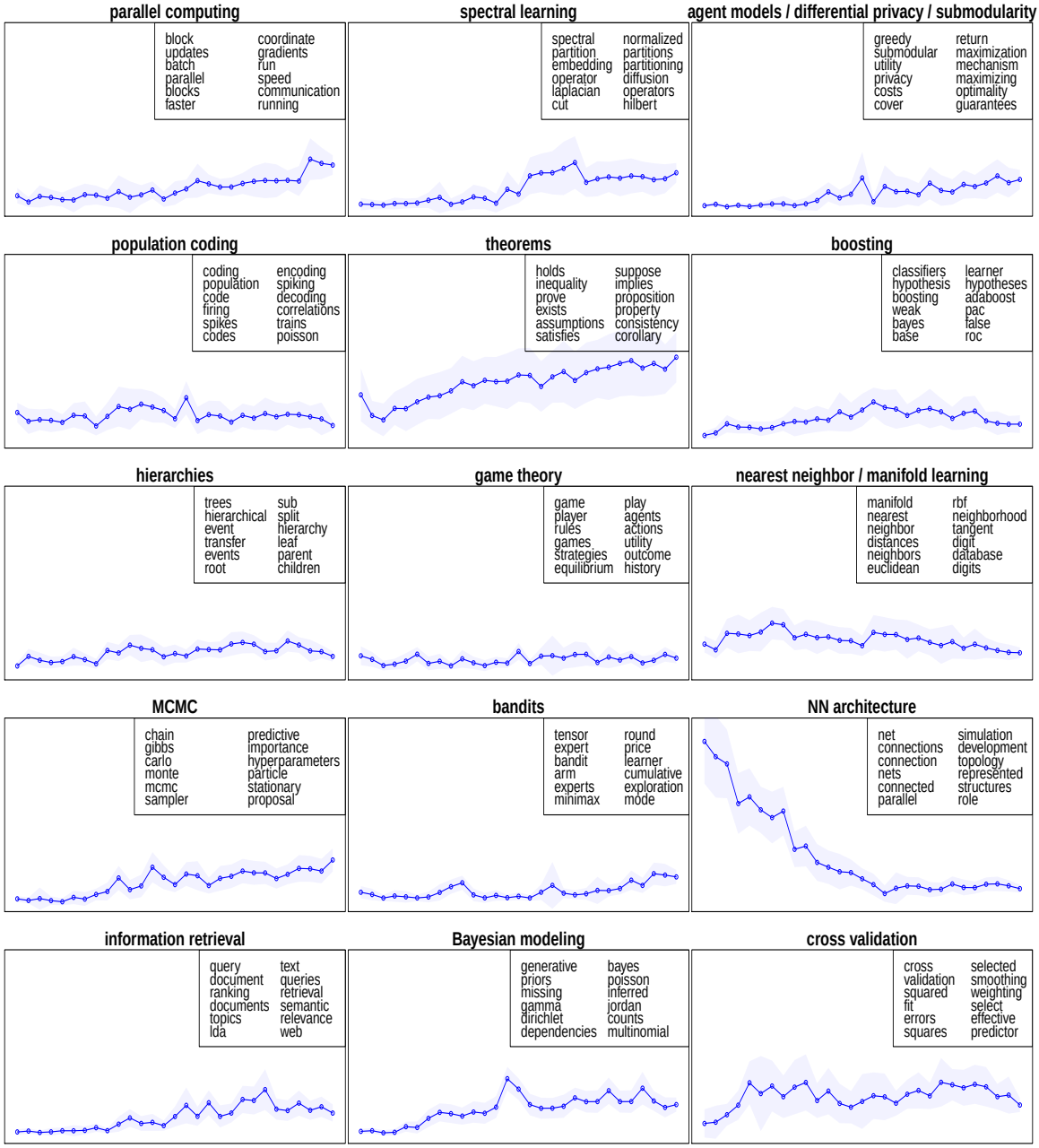


Figure 13: Posterior topic proportions over the years 1987-2015 and 12 most likely words for each topic (NIPS data set).

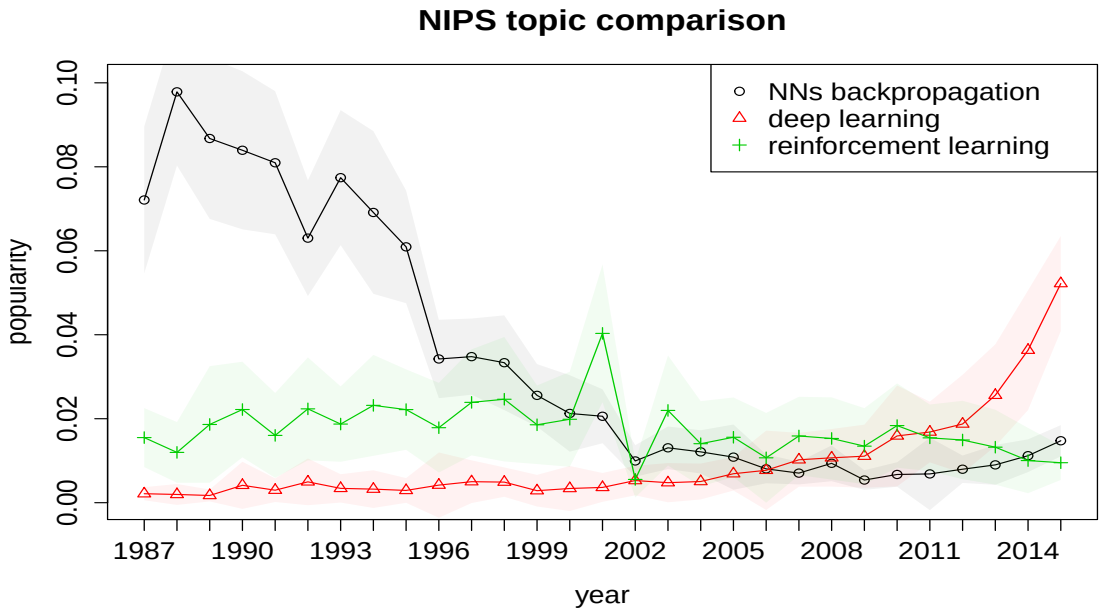


Figure 14: Comparison of a set of posterior topic proportions over the years 1987-2015 (NIPS data set).

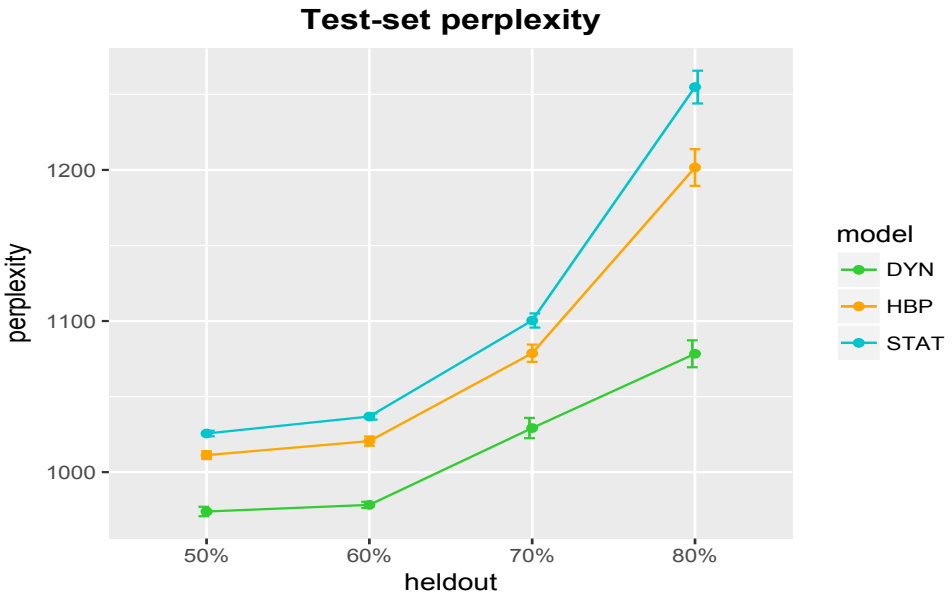


Figure 15: Comparison between the test-set perplexity of our dynamic model (DYN) against its hierarchical (HBP) and static (STAT) counterparts after holding out different percentages of words (NIPS data set). The dynamic model consistently outperforms the other two models, with a substantial difference when the percentage of held-out words is large.

8. Conclusion

We have presented a new framework for generating dependent IBPs by means of a novel time-evolving beta process, whereby feature probabilities evolve over time and are marginally distributed as in the beta process. At each time point items are exchangeable, and the two-parameter IBP is recovered. The key insight has been building on the PRF from population genetics to derive a suitable model for the prevalence and evolution of features over continuous time. We have developed an interesting MCMC framework for exact posterior inference with this model, and presented an alternative finite-dimensional approximation where the number of features is fixed.

As an application of the WF-IBP, we have described a time-dependent focused topic model that builds on Williamson et al. (2010b). The WF-IBP topic model allows for a flexible evolution of the popularity of an unknown number of topics over time, and compares favorably to HDP models by decoupling topic probabilities and within-document topic proportions. We have used our model to explore the data set consisting of the full text of NIPS conference papers from 1987 to 2015 and obtained an interesting visualization of how the popularity of the underlying topics evolved over these 29 years. In addition, test-set perplexity results have shown that incorporating time also improves on the predictive performance of the model.

A number of directions for future work are open. First, as $K \to \infty$ the fixed-K approximation marginally converges to the infinite model, and simulations showed that their dynamics are remarkably similar; further work could formally investigate the exact relationship between the two dynamics. Second, the current MCMC framework could be generalized to include inference on the IBP parameters and the W-F diffusion time step. Third, an extension of this work could modify the PRF by letting features evolve according to a more general W-F diffusion with selection and recombination, which would allow for feature-specific drifts in popularity and the coupled evolution of different features, respectively. Finally, our novel time-dependent beta process is a general construction with applications not limited to topic modeling. Different data and likelihood models could be explored following our work, for applications such as the modeling of time-evolving social networks or gene expression patterns.

Acknowledgments

The authors would like to thank the anonymous reviewers and the editor for their insightful and constructive feedback. Valerio Perrone is supported by EPSRC [EP/L016710/1]. Paul Jenkins is supported in part by EPSRC [EP/L018497/1]. Dario Spanò is supported in part by CRiSM, an EPSRC-HEFCE UK grant. Yee Whye Teh is supported by EPSRC for research funding under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 617071.

Appendix A. Proof of Theorem 1

The mean density of a PRF with immigration parameter α is

$$l(x) = \alpha m(x)dx,$$

where

$$m(x) = \frac{e^{I(x)}}{\sigma^2(x)}, \quad \text{with } I(x) := \int_0^x \frac{2\gamma(y)}{\sigma^2(y)}dy, \tag{12}$$

is the speed density of the process (see Griffiths, 2003). γ and σ are the drift and diffusion terms as defined in (2) and (3). Plugging (2) with $\mu = 0$ and (3) into the integral, we have

$$I(x) = \beta \ln(1 - x)$$

so that

$$m(x) = \frac{(1-x)^\beta}{x(1-x)} = x^{-1}(1-x)^{\beta-1}.$$

It follows that

$$l(x) = \alpha m(x)dx = \alpha x^{-1}(1-x)^{\beta-1}dx$$

is the resulting mean density, which completes the proof.

Remark 2 *When $\mu, \beta = 0$ it is necessary to condition each diffusion on hitting the boundary 0 before 1 in reverse time, which leads to an extra term in the analogous result in Sawyer and Hartl (1992).*

Appendix B. Simulating from the model

Consider the problem of simulating from the model, namely generating feature probabilities and the corresponding feature allocation matrices at a discrete set of time points.

Simulating X Set a truncation level $u > 0$ and consider the task of simulating features whose probability is above the threshold u at two times t_0 and t_1. As we know how to simulate marginally from the beta process, we can first generate the feature probabilities above u at time t_0 and let them evolve independently to time t_1. This yields features whose probability is greater than u at time t_0, meaning that we are still missing those features whose probability is below u at time t_0. To simulated these, we proceed as follows: we generate these features by drawing them from the beta process at time t_1 and propagate them backwards to time t_0. Finally, in order not to double-count features, all features that in the reverse simulation have probability greater than u at time t_0 have to be rejected.

Now translate these ideas into the following sampling scheme. At time t_0, sample from a truncated version of the PRF, namely from a Poisson process on $[u, 1)$ with rate measure $\alpha x^{-1}(1-x)^{\beta-1}dx$. This can be done, for instance, via an adaptive thinning scheme as described in Ogata (1981). As the truncation level u eliminates the point zero which has an infinite mass, this sampling procedure yields an almost surely finite number of samples. Denote by $\mathcal{K}$ the resulting set of feature indices and proceed as follows.

1. For all $k \in \mathcal{K}$, simulate $X_k(t) \mid \{X_k(t_0) = x_k(t_0)\} \sim \mathrm{WF}(0, \beta)$ for $t \in [t_0, t_1]$.

2. At time t_1, sample the candidate newborn features $X(t_1)$ from the truncated PRF as above. Let $\mathcal{L}$ denote the resulting set of features.

3. For all $l \in \mathcal{L}$, simulate $X_l(t) \mid \{X_l(t_1) = x_l(t_1)\} \sim \mathrm{WF}(0, \beta)$ backwards for $t \in [t_1, t_0]$ and remove from $\mathcal{L}$ the indices in the set $\{l : x_l(t_0) \geq u\}$.

4. Generate $Z_{ikt} \mid \{X_k(t) = x_k(t)\} \overset{iid}{\sim} \mathrm{Bernoulli}(x_k(t))$, for $t = t_0, t_1$, $\forall k \in \mathcal{K} \cup \mathcal{L}$ and $\forall i = 1, \ldots, N_t$.

As mentioned previously, the idea behind steps 2 and 3 is to compensate for the features with probability smaller than u that were discarded when generating features at time t_0.

This construction generalizes to a set of time points $t = t_0, \ldots, t_T$, observing that at a given time $t_{t^*} \in \{t_1, \ldots, t_T\}$ step 3 needs to be modified by simulating the W-F diffusions backwards to time t_0 and removing from $\mathcal{L}$ the indices such that $\exists t \in \{t_0, \ldots, t_{t^*-1}\}$ such that $x_l(t) \geq u$.

Simulating Z and the underlying X Consider now the more complex task of simulating both the feature allocation matrices Z and the features X appearing in them. First note that, although the PRF describes the evolution of an infinite number of features, we can sample the feature allocation matrices Z_{t_0} and Z_{t_1} at two times t_0 and t_1 exactly, as a property of the IBP is that the number of observed features is almost surely finite (Griffiths and Ghahramani, 2011). It is then possible to sample the features that are active in at least one object at times t_0 and t_1 and the corresponding allocation matrices Z_{t_0} and Z_{t_1} as follows. First, draw Z_{t_0} from the IBP, and use its realisation to draw the posterior probabilities $X(t_0)$ of the features seen in Z_{t_0} as in the posterior beta process. Then, simulate from the W-F diffusion to propagate these features to time t_1 and generate Z_{t_1} using these feature probabilities. We are now only missing the columns of Z_{t_1} corresponding to the features that were seen at time t_1 but not at time t_0. To add those columns, first draw a candidate $Z^C_{t_1}$ from the IBP and the corresponding feature probabilities; then, simulate these candidate features backwards to time t_0 and accept them with probability $(1 - X(t_0))^{N_{t_0}}$ to account for the fact that they were not seen at time t_0. The columns of $Z^C_{t_1}$ corresponding to the rejected features are deleted. Translating these ideas into an algorithm, consider the following steps.

1. Draw $Z_{t_0} \sim \mathrm{IBP}(\alpha, \beta)$ and index the resulting columns as $1, \ldots, K_1$.

2. For $k = 1, \ldots, K_1$ draw the corresponding feature probabilities

$$X_k(t_0) \mid \{Z_{t_0} = z_{t_0}\} \sim \mathrm{Beta}\left(n_{kt_0}, \beta + N_{t_0} - n_{kt_0}\right).$$

3. For $k = 1, \ldots, K_1$ simulate $X_k(t) \mid \{X_k(t_0) = x_k(t_0)\} \sim \mathrm{WF}(0, \beta)$ for $t \in [t_0, t_1]$ and set $X_k(t_1) = x_k(t_1)$.

4. Sample $Z^1_{t_1}$ by drawing each component $Z^1_{ikt_1} \mid \{X(t_1) = x(t_1)\} \overset{iid}{\sim} \mathrm{Bernoulli}(x_k(t_1))$, where $k = 1, \ldots, K_1$ and $i = 1, \ldots, N_{t_1}$.

Then, to sample the features that are active only at time t_1, add the following steps.

5. Draw a candidate $Z^{\mathrm{C}}_{t_1} \sim \mathrm{IBP}(\alpha, \beta)$ and index the resulting columns as $K_1+1, \ldots, K_2$.
6. For $k = K_1 + 1, \ldots, K_2$ draw the corresponding candidate feature probabilities
$$X^{\mathrm{C}}_k(t_1) \mid \{Z^{\mathrm{C}}_{t_1} = z^{\mathrm{C}}_{t_1}\} \sim \mathrm{Beta}\left(n_{kt_1}, \beta + N_{t_1} - n_{kt_1}\right).$$
7. For $k = K_1 + 1, \ldots, K_2$, simulate $X_k(t) \mid \{X_k(t_1) = x_k(t_1)\} \sim \mathrm{WF}(0, \beta)$ backwards for $t \in [t_1, t_0]$ and set $X_k(t_0) = x_k(t_0)$.
8. Accept the candidate columns of $Z^C_{t_1}$ with probability $(1 - x^C_k(t_0))^{N_{t_0}}$ and let Z_{t_1} be the matrix obtained by the union of the columns of $Z^1_{t_1}$ with the accepted columns of $Z^C_{t_1}$.

Note that the rejection in the last step is a way to account for the fact that we are considering features that are active for the first time at time t_1. When considering a general set of time points $t_0, \ldots, t_T$, it is necessary to account for the features that are active for the first time at each time $t_1, \ldots, t_T$. In this more general case, features seen for the first time at time t_{t^*} need to be accepted with probability $\prod_{t=t_0}^{t_{t^*-1}} (1 - x^C_k(t))^{N_t}$, as they were not seen in any of the feature allocation matrices at the time points before t^*.

Appendix C. Derivation of full conditionals

As observed in Williamson et al. (2009), in this topic modeling setting there are two equivalent ways of generating documents. Either the total number of words is sampled from a negative binomial $\mathrm{NB}(\sum_k z_{ikt}\phi_{kt}, 1/2)$ and then the topic and word assignments are drawn, or the number of words generated by each topic is drawn from $\mathrm{NB}(z_{ikt}\phi_{kt}, 1/2)$ and then the word assignments are picked. Following Williamson et al. (2009) closely, we make use of the latter construction to derive the full conditional distributions for the Gibbs sampler of the WF-IBP topic model.

Full conditional of a_{il} Recall that $a_{ilt} = k$ indicates that the lth word in document i at time t is assigned to topic k. We have

$$\begin{aligned} p(a_{ilt} = k \mid a_{-il}, Z_t, w_{ilt}, \phi_t) &\propto p(w_{ilt} \mid a_{ilt} = k) p(a_{ilt} = k \mid a_{-il}, z_{ikt}, \phi_{kt}) \\ &\propto p(w_{ilt} \mid a_{ilt} = k)(n^i_{kt} + \phi_{kt} z_{ikt}), \end{aligned}$$

where the last step is given by integrating out θ_{it}, namely the distribution over topics in document i at time t, and using the Dirichlet-Categorical conjugacy. Recall that n^i_{kt} denotes the number of words assigned to topic k in document i at time t and that n_{kt} denotes the total number of words assigned to topic k at time t, both excluding the assignment a_{il}. Similarly, integrating out the parameter ρ_k representing the distribution over words of topic k and using the Dirichlet-Categorical conjugacy, we have that

$$p(w_{ilt} \mid a_{ilt} = k) = \frac{(n^{w_{il}}_k + \eta)}{n_k + \eta D - 1},$$

which, plugged into the previous equation, gives the desired full conditional.

Full conditional of ϕ_k and γ We have that

$$\begin{aligned} p(\phi_{kt}, \gamma \mid n_{kt}, X(t), Z_t) &\propto p(\phi_{kt}, \gamma, n_{kt}, X(t), Z_t) \\ &\propto p(\phi_{kt} \mid \gamma) P(\gamma) P(n_{kt} \mid Z_t, \phi_{kt}), \end{aligned} \tag{13}$$

where

$$p(n_{kt} \mid Z_t, \phi_{kt}) = \prod_{i=1}^{N_t} p(n_{kt}^i \mid Z_{ikt}, \phi_{kt}) = \prod_{i=1}^{N_t} \mathrm{NB}(n_{kt}^i; Z_{ikt}\phi_{kt}, 1/2).$$

Note that $p(\phi_{kt} \mid \gamma)$ is distributed according to its prior Gamma$(\gamma, 1)$ and γ according to a chosen hyper-prior. The result follows immediately by plugging these three distributions into (13).

Full conditional of Z_{ikt} Recall that n_{ikt} denotes the total number of words assigned to topic k in document i at time t. If $n_{ikt} > 0$, then the corresponding entry Z_{ikt} is active with probability 1. If $n_{ikt} = 0$, we have

$$\begin{aligned} &p(Z_{ikt} = 1 \mid Z_{-(ik)t}, n_{ikt} = 0, X_k(t), \phi_{kt}, S_t) = \\ &\frac{p(Z_{ikt} = 1, Z_{-(ik)t}, n_{ikt} = 0, X_k(t), \phi_{kt}, S_t)}{p(Z_{-(ik)t}, n_{ikt} = 0, X_k(t), \phi_{kt}, S_t)}. \end{aligned}$$

The numerator is equal to

$$\begin{aligned} p(n_{ikt} = 0 \mid Z_{ikt} = 1, \phi_{kt}) p(S_t \mid Z_{ikt} = 1, Z_{-(ik)t}) p(Z_{ikt} = 1 \mid X_k(t)) p(\phi_{kt}, X_k(t), Z_{-(ik)t}) = \\ \mathrm{NB}(0; \phi_{kt}, 1/2) \frac{1}{x_1^{min}(t)} x_k(t) p(\phi_{kt}, X_k(t), Z_{-(ik)t}) \end{aligned}$$

Denoting by C the product of all terms not depending on z_{ikt}, we have

$$p(Z_{ikt} = 1 \mid Z_{-(ik)t}, n_{ikt} = 0, X_k(t), \phi_{kt}, S_t) = C \frac{1}{2^{\phi_{kt}}} \frac{1}{x_1^{min}(t)} x_k(t). \tag{14}$$

By the same token, we have

$$p(Z_{ikt} = 0 \mid Z_{-(ik)t}, n_{ikt} = 0, X_k(t), \phi_{kt}, S_t) = C \frac{1}{x_0^{min}(t)} (1 - x_k(t)). \tag{15}$$

As the two probabilities must sum to 1, we have that

$$C = \frac{2^{\phi_{kt}} x_1^{min}(t) x_0^{min}(t)}{x_0^{min}(t) x_k(t) + 2^{\phi_{kt}} x_1^{min}(t)(1 - x_k(t))},$$

which, plugged into equations 14 and 15, gives the result.

References

A. Ahmed and E. P. Xing. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream. *UAI*, abs/1203.3463, 2012.

A. Amei and S. Sawyer. A time-dependent Poisson random field model for polymorphism within and between two related biological species. *Annals of Applied Probability*, 20(5): 1663–1696, 2010.

A. Amei and S. Sawyer. Statistical inference of selection and divergence from a time-dependent poisson random field model. *PLoS ONE*, 7(4):e34413, 2012.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistal Society B*, 72:269–342, 2010.

A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On Smoothing and Inference for Topic Models. *UAI*, pages 27–34, 2009.

D. M. Blei and J. D. Lafferty. Dynamic topic models. *ICML*, pages 113–120, 2006.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *JMLR*, 3:993–1022, 2003.

A. R. Boyko, S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez, K. E. Lohmueller, M. D. Adams, S. Schmidt, J. J. Sninsky, S. R. Sunyaev, T. J. White, R. Nielsen, A. G. Clark, and C. D. Bustamante. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, 4(5), 2008.

C. D. Bustamante, J. Wakeley, S. Sawyer, and D. L. Hartl. Directional Selection and the Site-Frequency Spectrum. *Genetics*, 159(4):1779–1788, 2001.

C. D. Bustamante, R. Nielsen, and D. L. Hartl. Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theoretical Population Biology*, 63(2):91–103, 2003. ISSN 0040-5809.

C. E. Dangerfield, D. Kay, S. MacNamara, and K. Burrage. A boundary preserving numerical algorithm for the Wright-Fisher model with mutation. *BIT Numerical Mathematics*, 52(2):283–304, 2012.

F. Doshi-Velez and Z. Ghahramani. Accelerated Sampling for the Indian Buffet Process. *ICML*, pages 273–280, 2009.

A. Dubey, A. Hefny, S. Williamson, and E. P. Xing. A Nonparametric Mixture Model for Topic Modeling over Time. *SDM*, pages 530–538, 2013.

S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*. Wiley series in probability and mathematical statistics. J. Wiley & Sons, 1986. ISBN 0-471-08186-8.

W. J. Ewens. *Mathematical Population Genetics*. Springer-Verlag, Berlin, 2004.

S. Gershman, P. I. Frazier, and D. M. Blei. Distance Dependent Infinite Latent Feature Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):334–345, 2015.

Z. Ghahramani, T. L. Griffiths, and P. Sollich. Bayesian nonparametric latent feature models. *Bayesian Statistics*, pages 201–225, 2007.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schlkopf, and A. J. Smola. A Kernel Method for the Two-Sample-Problem. *NIPS*, pages 513–520, 2006.

R. C. Griffiths. The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theoretical Population Biology*, 64(2):241–251, 2003.

T. L. Griffiths and Z. Ghahramani. The Indian Buffet Process: An Introduction and Review. *JMLR*, 12:1185–1224, 2011.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101:5228–5235, 2004.

R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet*, 5(10), 2009.

D. L. Hartl, E. N. Moriyama, and S. A. Sawyer. Selection intensity for codon bias. *Genetics*, pages 227–234, 1994.

P. A. Jenkins and D. Spanò. Exact simulation of the Wright-Fisher diffusion. *Annals of Applied Probability*, 2017. To appear.

S. Karlin and H. M. Taylor. *A Second Course in Stochastic Processes*. Academic Press, 1981.

J. F. C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1): 59–78, 1967.

Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

K. T. Miller, T. L. Griffiths, and M. I. Jordan. The Phylogenetic Indian Buffet Process: A Non-Exchangeable Nonparametric Prior for Latent Features. *UAI*, abs/1206.3279: 403–410, 2012.

Y. Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–30, 1981.

V. Rao and Y. W. Teh. Spatial Normalized Gamma Processes. *NIPS*, pages 1554–1562, 2009.

S. A. Sawyer and D. L. Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132(4):1161–1176, 1992.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Y. W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking Construction for the Indian Buffet Process. *AISTATS*, 11:556–563, 2007.

R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. *AISTATS*, 2:564–571, 2007.

S. G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1):45–54, 2007.

S. Williamson, C. Wang, K. Heller, and D. Blei. Focused Topic Models. *NIPS Workshop on Applications for Topic Models: Text and Beyond*, pages 1–4, 2009.

S. Williamson, P. Orbanz, and Z. Ghahramani. Dependent Indian Buffet Processes. *AISTATS*, pages 924–931, 2010a.

S. Williamson, C. Wang, K. A. Heller, and D. M. Blei. The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling. *ICML*, pages 1151–1158, 2010b.

S. H. Williamson, R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen, and C. D. Bustamante. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *PNAS*, 102(22):7882–7887, 2005.

M. Zhou, H. Yang, G. Sapiro, D. B. Dunson, and L. Carin. Dependent Hierarchical Beta Process for Image Interpolation and Denoising. *AISTATS*, 15:883–891, 2011.